de lege

Law, AI and Digitalisation

Editors: Katja de Vries and Mattias Dahlberg

JURIDISKA FAKULTETEN
I UPPSALA

de lege

Juridiska Fakulteten i Uppsala

Årsbok 2021

Redaktör för skriftserien
Mattias Dahlberg

Law, AI and Digitalisation

Eds. Katja de Vries
and Mattias Dahlberg

§
iUSTUS
FÖRLAG

The Emil Heijne's Foundation for Research in Legal Science (Emil Heijnes Stiftelse för rättsvetenskaplig forskning) has generously provided financial support for the printing of this yearbook.

Upplaga 1:1
ISSN 1102-3317
ISBN 978-91-7737-167-0
Produktion: eddy.se ab, Visby 2022
Omslag: John Persson
Förlagets adress: Box 1994, 751 49 Uppsala
Telefon: 018-65 03 30
Webbadress: www.iustus.se, e-post: kundtjanst@iustus.se

Printed by Dimograf, Poland 2022

Preface

This volume of *De lege* includes articles by scholars on a wide range of subjects, all dealing with "artificial intelligence" (AI) and the digitalisation of society. Some of the authors have dealt with AI and digitalisation for a long time, others are specialists in other fields of law and have, in this volume, turned their focus on these matters. We are especially glad to have contributors from other faculties within Uppsala University, as well as authors from other universities.

We wish to thank all the contributing authors for the co-operation in making this volume possible. We also want to thank Vice Dean, Professor Anna Jonsson Cornell for support during the project. The English proofreading has generously been funded by Professor Bengt Domeij and his Huselius funds. The printing of this volume has been funded by the Emil Heijne's Foundation for Research in Legal Science (*Emil Heijnes stiftelse för rättsvetenskaplig forskning*), for which we are grateful.

Uppsala, February 2022

Katja de Vries	*Mattias Dahlberg*
Editor	Editor

Contents

PART I

AI, DIGITALISATION AND LAW: FOUNDATIONAL EXPLORATIONS

PART II

CHALLENGES POSED BY AI AND DIGITALISATION TO PARTICULAR FIELDS OF LAW

PART III

AI AND DIGITALISATION IN PRACTICE: LEGAL PERSPECTIVES

Katja de Vries

Introduction to the De lege Yearbook 2021: Law, AI and Digitalisation

1 AI and Digitalisation: what is it?

Artificial Intelligence (AI) is everywhere – but *what* is AI? Those working with AI often discard the notion as a vague buzzword and instead prefer speaking about Machine Learning (ML), which is the dominant set of techniques driving the current AI revolution. ML, in turn, has deep ties to good old-fashioned statistical methods. As a well-known joke states: *When you're fundraising, it's AI. When you're hiring, it's ML. When you're implementing, it's statistics*. No wonder that the definition of AI is a contested topic in the proposed AI Act.[1] In the initial version, proposed by the Commission on the 21st of April 2021, it says in Article 3(1):

> 'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I[2] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with

[1] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final), 21 April 2021.

[2] Annex I states: *Artificial Intelligence techniques and approaches referred to in Article 3, point 1*:

(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

Criticised for being a too broad and vague grouping of techniques, the Council proposed to reformulate this in the presidency compromise text[3] (29th of November 2021) into a much narrower definition, which mainly comprises ML systems that transform input data (such as passenger data) into output data (such as "potential terrorist" and "ordinary traveller") according to a rule (that is, a model) that is bottom-up inferred from training data (that is, existing data relating to confirmed terrorists and ordinary travellers, respectively):

> 'artificial intelligence system' (AI system) means a system that (i) receives machine and/or human-based data and inputs, (ii) infers how to achieve a given set of human-defined objectives using learning, reasoning or modelling implemented with the techniques and approaches listed in Annex I[4], and (iii) generates outputs in the form of content (generative AI systems), predictions, recommendations or decisions, which influence the environments it interacts with;

The definition of "AI system" will no doubt be further contested and refined during the remainder of the legislative process. From a computer science perspective, AI is as vague a notion as it ever was. Yet, discussing AI has become unavoidable, precisely *because* it is currently crystallizing as a concept within legal, political, policy and public discourse and as an object for legal and ethical regulation. Law, like always, is pragmatic. It does not aim to pinpoint a metaphysical essence of an AI system. Instead, it defines AI in a way that fits the regulatory purpose. For example, in the aforementioned definitions of 'AI system' from the proposed AI Act, the last part reads that its outputs 'influence the environments it interacts with'. Given that the AI Act aims to regulate AI systems that could adversely impact on society or individuals, and explicitly excludes

(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) Statistical approaches, Bayesian estimation, search and optimization methods.

[3] Council of the European Union, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – Presidency compromise text, 29 November 2021, available at: https://www.statewatch.org/media/2963/eu-council-ai-act-compromise-text-14278-21.pdf.

[4] Annex I (n. 2).

any general-purpose AI from its scope,[5] this definition makes sense from a regulatory perspective. Yet, if one was to create a dictionary entry for "AI", an exclusion of non-influential, general-purpose AI would be an unwarranted limitation of the term's scope.

It is often a conglomerate of techniques, applications or situations that is the object of regulation, not merely AI in a narrow technical sense. As Bruno Debaenst (Chapter 1) poignantly argues, the impact of the current AI revolution, that is, the fourth industrial revolution, on law, cannot be understood in isolation from earlier industrial revolutions. While some legal challenges posed by the AI revolution are truly novel, other challenges overlap with those raised by the third industrial revolution, that is, the digital revolution of computer automation. To underline this regulatory intertwinement, this book is entitled *Law, AI and Digitalisation*: addressing regulatory digitalisation challenges is often a necessary step before advancing to challenges posed by AI as such. An example is intellectual property (IP) questions, raised by avatars mimicking human behaviour, looks or speech that are created using AI techniques, so-called *synthetic* avatars. Such avatars can, for example, be used as *vocational* avatars, that is, as digital stand-ins for human professionals, such as teachers[6] or actors.[7] Vocational avatars can be created by training them on footage from flesh and blood equivalents, which involves making copies of that footage for training purposes. Anyone who wants to create vocational avatars will, thus, have to find training materials and tackle the question of who owns the copyright over them. These are questions that are already highly relevant in today's world where teaching has become increasingly digitised due to the pandemic. Can a university re-use a recorded lecture if a teacher does not consent? Who owns the copyright over digitised teaching materials produced at today's universities – the university teacher or the employer? And if it is the university teacher who holds the copyright, does the university have a right of usage? As Marianne Rødvei Aagaard (Chapter 10) argues, in Sweden, teachers are normally copyright holders of their teaching materials, but the university could, in a limited set of situations, also have some usage rights. While Rødvei Aagaard

[5] Article 52a and Recital 70a AI Act (n. 3).

[6] Kristin Houser (2018). The world's first digital teacher just debuted in New Zealand. Tomorrow's teachers won't all be flesh and blood, The Byte, online available at: https://futurism.com/the-byte/digital-teacher-new-zealand-will.

[7] Mathilde Pavis (2021). Rebalancing our regulatory response to Deepfakes with performers' rights. Convergence, 27(4), 974-998.

does not discuss the reuse of teaching material as training material for a robot teacher, the question of the permissibility of such reuse is likely to develop along the same lines as more conventional forms of reuse. Even if the conclusion might be different (for example, because re-streaming a recorded lecture was an intended and foreseeable re-use, while training an avatar-teacher was not), the analysis of the avatar situation would still make use of the same legal framework and concepts.

The reuse of digital teaching materials is a clear example of continuity between legal challenges raised by the third industrial revolution (computer automation and digitalization) and the fourth industrial revolution (AI). In contrast, certain AI systems, such as the high frequency trading algorithms used in financial markets (Magnus Strand, Annina H. Persson, and Malou Larsson Klevhill, Chapter 22) or the AI-based automated grading systems, discussed by Cecilia Magnusson Sjöberg and Rebecka Weegar (Chapter 16) clearly go beyond the legal challenges raised by computer automation and digitalisation. Automated decision-making (ADM) and risk assessment systems, which are discussed by several authors in this volume, can be based on ML techniques, and more traditional systems, based on digitisation and automation. The choice of topics in this volume is characterised by a legal pragmatism: it does not only address AI in a narrow sense, but it also tackles more classical digitalisation and computer automation questions in as far as they continue to be relevant in an AI context.

2 AI and Digitalisation: where is it?

While the question of *what* AI is might be contested, the question of *where* it is can be answered quite easily: it is almost everywhere. During the last few years, AI systems have become so ubiquitously applied that they hardly left any aspect of life untouched. AI is used in surveillance, healthcare, smart cars, decisions relating to social welfare benefits, financial investments, generation of texts and other media, grading, research, warfare, fraud detection, border control, etc., etc. The list could easily be extended to fill the rest of the page. With every field that AI touches upon, it also raises a particular set of legal questions. These questions are often very fundamental questions: when activities, which used to be a matter of human discernment – fighting a war, deciding how to invest money, making a medical diagnosis or driving a car – are delegated to AI, it forces us to revisit the ground rules of these activities.

The contributions in this volume represent the multiplicity, broadness and richness of the legal questions raised by AI. They are organised into three parts: (I) AI, digitalisation and law: foundational explorations, (II) challenges posed by AI and digitalisation to particular areas of law, and (III) AI and digitalisation in practice: legal perspectives.

3 AI, digitalisation and law: foundational explorations (Chapters 1–7)

The first part of this volume (Chapters 1–7) discusses legal foundational questions raised by AI. How to deal with a new phenomenon? In Stanley Kubrick's movie, *Space Odyssee 2001*, bewildered apes gather around a rectangular black monolith that lands on the prehistoric earth, and as their fear makes place for curiosity, they begin to touch and explore the surface of the unknown object. AI, like Kubrick's black monolith, is a new phenomenon that needs to be approached with explorative questions. How does the impact of AI compare to earlier technological revolutions? How should we relate to AI? How do we steer it? What data do we feed it? The chapters gathered in *Part I*: *AI and Law – foundational explorations* all breathe an explorative spirit that is similar to the one that dominates the beautiful opening sequence of *Space Odyssee 2001*. In Chapter 1, *Bruno Debaenst* moves between legal history and legal futurology. He looks at the current AI revolution (the fourth industrial revolution) and compares its potential impact on law with the first, second and third industrial revolution. Chapters 2, 3 and 4 look into ways to relate to AI and the legal status granted to it. In Chapter 2, *Annika Waern* uses her field of expertise, Human-Computer Interaction, to present the different types of relations that humans can have with AI; she argues that not all types of relations are equally desirable from a societal and normative perspective. In Chapter 3, *Bert Lehrberg* discusses the question if AI could be attributed some form of legal personhood. He connects this question to three settings: algorithmic contracting, autonomous functioning vehicles and artificial general intelligence. In Chapter 4, *Anni Carlsson* addresses the question of legal personality for AI from a completely different angle; she looks for legal guidance in fiction, as 'law is also present in an imaginary society'.[8] She shows how humanoid robots can challenge the legal dichotomy of legal property and personhood. In Chapter 5, *Stanley Greenstein, Panagiotis Papapetrou and Rami Mochaourab* discuss the difficulties in building human values into the design of AI when the promoting of one human value is often at the expense of another competing human value. After a theoretical discussion of value sensitive design, the authors show how three values derived from the field of data protection, namely explainability, privacy and accuracy, are used for the construction of an AI system that predicts the diagnosis of future patients and that is trained on medical patient data. Training data is an indispensable basic ingredient for the creation of any AI model. The last two Chapters (6–7) in the first part of this volume deal with the question of where to find (training) *data* to fuel AI innovation. In Chapter 6, *Katja de Vries* discusses the options for a researcher who wants to use personal data. Is it better to use personal data (which entails compliance with data protection legislation) or is it possible to use non-personal surrogates, such as synthetic data or data of deceased people? In Chapter 7, *Bengt Domeij* throws light on some newly proposed pieces of EU legislation, such as the Digital Markets Act and the Data Act. He dives into the question of how the EU proposes to facilitate business-to-business data sharing and under which exceptional circumstances businesses could be legally forced to share industrial data.

In thinking about specific legal challenges following from AI and digitalisation, one could, broadly speaking, take one of the following two approaches: either one departs from a particular *field of law* and discusses the various challenges posed to it, or one takes the opposite direction by departing from AI and digitalisation in practice, that is, a particular *tool or field of application*, and then studying which legal challenges are evoked. This division is the organisational principle that informs the second and third part of this volume. It should, however, be underlined that the division is not watertight. Most contributions in this volume look at one or more legal fields, as well as at a particular AI application. Yet, in most contributions, one of these perspectives is more dominant. The *second part* of this volume (Chapter 8–14) gathers contributions where the dominant perspective is an engagement with the challenges posed by AI to particular legal fields, such as constitutional law, intellectual property law or international humanitarian law, whereas the contribu-

[8] Jaakko Husa, 'Comparative law, literature and imagination: Transplanting law into works of fiction' (2021). 28(3) Maastricht J Eur Comp Law 371, 383.

tions gathered in the *third part* (Chapters 15–22) are characterised by predominantly departing from AI and digitalisation in practice, that is, a particular tool or application.

4 Challenges posed by AI and digitalisation to particular fields of law (Chapters 8–14)

In Chapter 8, *Markku Suksi* provides a thorough discussion of the use of ADM tools in relation to Finnish *constitutional and administrative law*, notably in relation to central concepts such as legality and rule of law, and makes some interesting comparisons to the Swedish legal framework. In Chapter 9, *Inger Österdahl* looks at the possibilities for regulating lethal autonomous weapon systems in the context of *the law in war*, or *international humanitarian law*, notably discussing the framework provided by the United Nations (UN), the Swedish policy position, and the controversial role of human control over advanced weapon systems. The authors of the following two chapters (10–11) all engage with AI and digitisation through the lens of *intellectual property law*. In Chapter 10, *Marianne Rødvei Aagaard* addresses a question that is of great practical importance in a time where the pandemic has made teaching increasingly digital: namely, who owns the copyright over digitised teaching materials – the university teacher or the employer? In Chapter 11, *Silvia Carretta* lays out the problems associated with content moderation of copyright-infringing material on the internet and the controversial Article 17 of the Copyright Directive 2019/790, which establishes a platform liability which, in practice, seems to make the use of automated upload-filters unavoidable. In Chapter 12, *Mikael Hansson* takes a *labour law* perspective. Can an AI system be an employer and hence have employer liabilities? Hansson discusses this thought-provoking question by looking at two Swedish labour law cases and one case decided upon by the Court of Justice of the European Union (CJEU) and concludes that AI cannot be held liable in the way a human employer can. In Chapter 13, *Vladimir Bastidas* looks at the practice of algorithmic personalised price-discrimination from an *EU competition law* perspective. When personalised pricing entails that an undertaking, to a certain extent, may determine market conditions, could this result in the applicability of Article 102 Treaty of the Functioning of the European Union (TFEU), which prohibits abuse by an undertaking in a dominant position and sets limits for the exercise of market power. Through a detailed analysis of case law, Bastidas raises several pivotal questions, for example, if the abuse in Article 102 TFEU only refers to behaviour directed towards other competitors or if it could also include the differential treatment of end-consumers. All in all, Bastidas concludes that, in theory, a wide interpretation of Article 102 TFEU may include cases on personalised prices, even though such an interpretation does not seem very likely given the existing case law. Part 2 of this volume is concluded by a contribution by *Mattias Dahlberg* that departs from the field of *tax law*. In Chapter 14, Dahlberg discusses the problem that the digital economy requires hardly any physical presence in the market state and that giant tech companies can get away with paying very little taxes by establishments in low-tax jurisdictions, which are not representative of the markets that they serve. New legislation on the taxation of multinational enterprises proposes to break away from traditional taxation principles and allow market states to tax the income generated in that state. However, to determine the income generated in a state could require extensive surveillance of how consumers use digital services and could result in infringements on consumer privacy.

5 AI and digitalisation in practice: legal perspectives (Chapters 15–22)

Part 3 opens with two chapters looking at the legal implications of using *AI in educational settings*. In Chapter 15, *Liane Colonna* discusses the use of AI-tools in higher education (HE), such as remote-based proctoring or predictive learning analytics, from the perspective of the proposed AI Act.[9] Colonna questions how well the risk-based approach of the AI Act, which categorises AI systems as entailing unacceptable, high, limited or minimal risks, fits the reality of educational AI, where it will often be difficult to categorise the AI system because of the unforeseen uses that can emerge in complex and large HE institutions. Colonna also observes that a university, in many situations, is likely to be both the provider and the user of an AI system. Two other sources of concern are that the AI Act has a largely technocratic approach which puts a lot of responsibility on the risk self-assessment by the AI developer that seems to exclude a participatory discussion with universities, students and teachers, and that

[9] AI Act (n. 1).

the AI Act is unlikely to sufficiently mitigate the potential abuse caused by biometric technologies like facial recognition in educational settings. In Chapter 16, *Cecilia Magnusson Sjöberg and Rebecka Weegar* look at AI-based automated grading systems, both from a computer science and a legal perspective. They describe a pilot project at Stockholm University where a compulsory one-page written assignment was graded using an AI-model. The model was trained on assignments that were graded by human graders. The AI-assigned grades were compared with grades given by two human graders. Sjöberg and Weeberg discuss how the combination of AI and human grading could potentially lead to enhanced equal treatment of students.

The next two Chapters look at the use of *AI in healthcare*. In Chapter 17, *Santa Slokenberga* looks at three types of EU regulatory responses to medical ML used in paediatric care: the General Data Protection Regulation,[10] the proposed AI Act[11] and the Medical Device Regulation.[12] She argues that none of the surveyed legal instruments contribute to furthering the development and availability of the devices directly and thus, the EU misses a chance to contribute to reducing the therapeutic gap in paediatric medical care. In Chapter 18, *Charlotte Högberg and Stefan Larsson* ask what role transparency and explainability of AI could have in relation to patients' rights and information flows in Swedish health care. They argue that, as is often falsely argued, the highest quality of health care is not opposed to transparency and that, in fact, the best possible health care cannot be achieved without transparency.

In Chapter 19, *Katarina Fast* looks at the use of *AI in child-related social services*. She gives an overview of several AI tools that have been used by social services to identify children at high risk of maltreatment. Most of these tools have been adopted in a context of struggles related to increases in caseloads, funding cuts and staff shortages and ambitions to increase digitalisation. Due to legal, ethical and public trust problems,

[10] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and the repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) OJ L 119, 4.5.2016, 1–88.

[11] AI Act (n. 1).

[12] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and the repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance) OJ L 117, 5.5.2017, 1–175.

many of these tools have been discontinued. Fast, then, raises the question of what would be needed to make such tools compatible with relevant children's rights as laid down in the UN Convention of the Rights of the Child, and other legislations, such as the European Convention of Human Rights, the EU Charter of Fundamental Rights, General Data Protection Regulation 2016/679 and the proposed AI Act.

The following two Chapters look at *AI-based decision-making and governance in the public sector*. In Chapter 20, *Stefan Larsson and Jonas Ledendal* present a comprehensive overview of the policy positions taken by the government and public authorities in Sweden towards the use of AI in the public sector. They compare the Swedish position with policies and legislative initiatives at the EU level, as well as with research on the use of AI in the public sector. One of their conclusions is the importance of creating a more specific legal and policy framework that articulates how AI-based decision (support) systems can be used in the public sector in a way that is in accordance with principles of good administration and the values steering government employees in Sweden ("*statliga värdegrunden*") and that preserves trust in public administration. In Chapter 21, *Johan Eddebo and Anna-Sara Lind* provide a lively encounter between a philosopher of religion and a legal scholar, discussing the question of if and how information, which affects discourses and opinion formation in the public sphere in potentially undesirable ways, should be governed by automated means. Eddebo and Lind introduce the notion of double intransparency, in relation to the influence exercised through algorithms within the framework of digital communication platforms and contemporary media technologies: it pertains to both our inability to access the algorithms as such, as well as the difficulties in reproducing and examining their actual effects. They discuss several possible mitigation strategies, such as the ones in the AI Act, and call for a deliberation that is both multidisciplinary and inclusive of all parts of civil society to consolidate the foundations of liberal democracy in an age of algorithmic content moderation.

The final contribution to this volume looks at the use of *AI in algorithmic financial trading*. In Chapter 22, *Malou Larsson Klevhill*, *Annina H. Persson and Magnus Strand* present an overview of the extremely complex national Swedish and EU legislative framework, regulating algorithmic high-frequency trading. One of the regulatory gaps identified by Klevhill, Persson and Strand is the lack of a clear civil liability regime: while administrative and criminal law sanctions can be good tools to promote legal compliance in the financial sector, the investor who suffers damages following a mistake in a financial trading algorithm is not helped by that.

6 Concluding thoughts: Law in a world organised by ML perceptions

As I have argued elsewhere,[13] one pivotal characteristic of ML (and thus of AI) in comparison to other technologies is the fact that it often revolves around *pattern* discovery, or, to put it in more anthropomorphic terms, it is a *meaning-attributing* technology. It perceives something *as* something. In order not to crash, a smart car needs to see a tree as a tree, a pedestrian as a pedestrian, a traffic sign as a traffic sign; an autonomous weapon system needs to distinguish foe from friends and civilians from military combatants; a financial investment AI needs to recognise a profitable investment opportunity from a bad one; a medical diagnostic AI tool should not mistake a malignant tumour and a benign one; a predictive model used by social services should correctly identify which child is at risk of being abused and which is not. This is the fundamental difference between a more conventional one like a knife and an AI-tool: the knife can cut a loaf of bread, but it does not *perceive* the bread as something (healthy or unhealthy, old or fresh, etc.). AI-tools help humans to make sense of the world: to see patterns and sometimes to propose new patterns: for example, one can imagine a kitchen-AI that not only recognises meals and gives the underlying recipe, but also suggests other recipes that might result in even better meals. Making sense of the world used to be a capacity belonging exclusively to humans, or, at the very least, to living beings: 'On an insignificant background of reality, imagination designs and embroiders novel patterns: a medley of memories, experiences, free fancies, absurdities and improvisations'.[14] What is so revolutionary about AI in the current AI-revolution is that it has introduced a category of

[13] Katja de Vries, Privacy, due process and the computational turn. A parable and a first analysis. In M. Hildebrandt & K. De Vries (eds.), *Privacy, Due Process and the Computational Turn. The Philosophy of Law Meets the Philosophy of Technology* (Routledge, 2013) 9–38.

[14] August Strindberg, The Dream Play (transl. Edwin Björkman; Charles Scribner's Sons, 1912), online available at: https://www.gutenberg.org/files/45375/45375-h/45375-h.htm In the Swedish original: '… på en obetydlig verklighetsgrund spinner inbillningen ut och väfver nya mönster: en blandning af minnen, upplefvelser, fria påhitt, orimligheter

pattern-recognising *tools* to support, or replace, humans in making sense of the world.

How does law regulate AI? How does law capture AI inside existing and new legal frameworks and concepts? The contributions in this volume show that this cannot be answered univocally, and that there are many answers. With law, the devil is always in the details. Law never operates through big ethical abstractions. Instead, it takes AI tools as they come: intertwined with other techniques and technologies, ranging from very simple automated systems to highly autonomous complex AI, mixing old legal questions with novel ones, strongly situated in practice with their own particularities, etc. This volume is, thus, a tribute to law's ways of dealing with AI in all its diversity and situatedness.

och improvisationer'. August Strindberg, Ett Drömspel in *Kronbruden. Svanehvit. Drömspelet*. (Stockholm, Iduns, 1902) 12.

Bruno Debaenst

The Digital Revolution from a Legal Historical Perspective[1]

1 Introduction

'*Still valid today is the lesson from the first industrial revolution – that the extent to which society embraces technological innovation is a major determinant of progress*'. The quote stems from Karl Schwab, the director of the World Economic Forum, who introduced the concept of the "*Fourth Industrial Revolution*" to a broader audience by writing a booklet on the topic in 2016.[2] One of his goals was '*to increase awareness of the comprehensiveness and speed of the technological revolution and its multifaceted impact*'.[3] This edition of *De Lege* on Artificial Intelligence and Law illustrates that awareness is undoubtedly growing.

Schwab's quote is also interesting because he makes a reference to the First Industrial Revolution. In uncertain times, people tend to look at the past to take lessons for the future. There have been three industrial revolutions before: periods of fundamental industrial and societal transformations that have changed the world forever. In this contribution, I will take a closer look at how law has been dealing with these earlier industrial revolutions. Can we learn something from the past? Are we just repeating a pattern or is something new going on? Which lessons can

[1] I would like to thank Katja de Vries, Rawaz Shanagar, Miljana Todorović, Sebastiaan Vandenbogaerde and Matthias Van Der Haegen for their valuable comments on this contribution.

[2] Karl Schwab, *The Fourth Industrial Revolution* (World Economic Forum 2016) 13.

[3] *Id.* 9.

we take from the past, to be better prepared for the future? Since long I have been fascinated by the industrial revolutions and their interaction with law. I did my doctoral research on this topic, by studying the process of juridification of workplace accidents in the nineteenth century in Belgium.[4] I am also increasingly interested in the ongoing digital revolution. A few years ago, I was surprised to find out that in the seminars on comparative legal history, my master students did not have a clue about what was going on. They literally had never heard of smart contracts, the Internet of Things or blockchain. How was this possible? Were we delivering already outdated lawyers to the labour market? This motivated me to include a lecture on "*legal futurology*" in our legal history course, to at least point out to the students the ongoing digital revolution and its many challenges and opportunities. In this text, I will try to bring the past and the future together.

2 The First Industrial Revolution – the Age of Steam

The First Industrial Revolution started in the mid-18th century in Great Britain. Coal mines fuelled steam engines that literally drove industrialisation. Textile factories mechanised. Already early on, the law was used to tame the industrial beast. In 1802, the Parliament of the United Kingdom passed the first of a series of Factory Acts, to protect workers in the mechanised cotton factories.[5] In France, Napoleon regulated the coal mines in 1810 and installed a mining inspection in 1813.[6] In 1831, Prussia dealt with its steam engines and in 1838 with the upcoming railroads, introducing a regime of strict liability.[7] Law also facilitated the indus-

[4] Bruno Debaenst, *Een proces van bloed, zweet en tranen. Juridisering van arbeidsongevallen in de negentiende eeuw in België* (KVAW 2011); see also Bruno Debaenst, 'A Study on Juridification. The Case of Industrial Accidents in Nineteenth Century Belgium' [2013] 81 (1-2) Legal History Review 247.

[5] E.P. Hennock, *The origins of the welfare state in England and Germany*, 1850–1914 (Cambridge University Press 2007) 73–85.

[6] Loi concernant les mines, les minières et les carriers 1810; Décret contenant les dispositions de police relatives à l'exploitation des mines 1813.

[7] Strict liability was motivated because of the dangerous nature of the railroad activities and covered personal injury, damage to transported goods and damage to any other goods, including damage to neighbouring land. Miquel Martín-Casals, 'Technological

trial revolution. The British patent system, for instance, was continually evolving and responding to the needs of the industrialising economy, without any legislative reform.[8] Inventors could easily obtain and enforce patent rights, which encouraged them to develop new technology. In the 1830s, several German states, such as Baden and Saxony, changed their expropriation legislation to facilitate the construction of railroads.[9] In other words, law played a pivotal role in regulating and facilitating the industrial revolution.[10] The other way around, the First Industrial Revolution seems not to have had any direct impact on the legal sector (legal education, legal professions, legal methodology).

3 The Second Industrial Revolution – the Age of Electricity

Around the 1870s, the Second Industrial Revolution began with new technologies, such as electricity, the telegraph and the telephone, chemical industry and the production line. Meanwhile, the sectors of the First Industrial Revolution continued to grow and to develop. Countries like Germany and Japan industrialised at a high speed. All around the world, railroads were constructed. Law increasingly had to deal with interesting legal questions raised by the ongoing industrialisation. One example was electricity, which was mostly invisible, powerful and had an incredible arrangement of new, fascinating possible applications. Lawyers had to figure out how they could legally frame this elusive matter. All kinds of

Change and the Development of Liability for Fault: A General Introduction', in: Miquel Martín-Casals (ed.), *The Development of Liability in relation to Technological Change* (Cambridge University Press 2010), 7. See also Nina vom Feld, *Staatsentlastung im Technikrecht. Dampfkesselgesetzgebung und überwachung im Preussen 1831-1914* (Klostermann 2007); M Eckardt, *Technischer Wandel und Rechtsevolution. Ein Beitrag zur ökonomischen Theorie der Rechtsentwicklung am Beispiel des deutschen Unfallschadensrecht im 19. Jahrhundert* (Mohr 2001). D. Ziegler, *Eisenbahnen und Staat im Zeitalter der Industrialisierung* (Steiner 1998).

[8] Sean Bottomley, *The British patent system during the industrial revolution*, 1700–1852 (Cambridge University Press 2014).

[9] Interestingly, the Swedish Expropriation Law of 1845 did not have a link with the construction of railroads. See Jonatan Bromander, *Expropriation i Sverige – en rättshistorisk analys* (examensarbete Juridiska institutionen Uppsala 2020).

[10] Marc Steinberg, *England's great transformation: law, labor, and the Industrial Revolution* (The University of Chicago Press 2016).

difficult legal issues emerged.[11] What was electricity? An object? Which legal qualification should it receive? Electricity could be produced, measured, traded, stolen, etc. Other difficulties arose. The Industrial Revolution developed organically; over the years, each country or even each company had developed its own standards. Lawyers and engineers gathered at international conferences to develop universal standards, for instance, regarding electricity.[12]

The Second Industrial Revolution increasingly fuelled the legal development. New branches of law popped up and old ones blossomed. "*Industrial law*" became the hot topic of the day. For instance, in Belgium in 1898, no less than three specialised journals on this topic saw the light of day.[13] Industrial law included *inter alia* patent law that had to deal with the numerous conflicts arising out of scientific discoveries and technical advances. Industrial workplace accidents also received increasing attention from lawyers, because of the many liability issues that arose. For instance, in 1884, a Belgian lawyer, Charles-Xavier Sainctelette, wrote a fascinating booklet titled "*De la garantie et de la responsabilité (accidents de transport et de travail)*" – "*On warrantee and liability (transport and workplace accidents)*".[14] Starting from railroad liability cases from the

[11] A famous example is the efforts by the German Reichsgereicht to define the railroad in 1879: '*[e]ine Eisenbahn […] ein Unternehmen [sei], gerichtet auf wiederholte Fortbewegung von Personen oder Sachen über nicht ganz unbedeutende Raumstrecken auf metallener Grundlage, welche durch ihre Konsistenz, Konstruktion und Glätte den Transport großer Gewichtmassen beziehungsweise die Erzielung einer verhältnismäßig bedeutenden Schnelligkeit der Transportbewegung zu ermöglichen bestimmt ist, und durch diese Eigenart in Verbindung mit den außerdem zur Erzeugung der Transportbewegung benutzten Naturkräften – Dampf, Elektrizität, tierischer oder menschlicher Muskeltätigkeit, bei geneigter Ebene der Bahn auch schon durch die eigene Schwere der Transportgefäße und deren Ladung usf. – bei dem Betriebe des Unternehmens auf derselben eine verhältnismäßig gewaltige, je nach den Umständen nur bezweckterweise nützliche oder auch Menschenleben vernichtende und menschliche Gesundheit verletzende Wirkung zu erzeugen fähig ist*'. One has to know what a railroad is to understand the definition. Roman Konertz & Raoul Schönhof, *Das technische Phänomen "Künstliche Intelligenz" im allgemeinen Zivilrecht* (Nomos 2020) 16.

[12] Miloš Vec, *Recht und Normierung in der Industriellen Revolution. Neue Strukturen der Normsetzung in Völkerrecht, staatlicher Gesetzgebung und gesellschaftlicher Selbstnormierung* (Klostermann Vittorio 2006).

[13] The *Revue pratique et juridique des accidents du travail* (1898–1899), the *Revue pratique du droit industriel* (1898–1919) and the *Revue des questions de droit industriel* (1898–1942).

[14] Charles-Xavier Sainctelette, *De la garantie et de la responsabilité (accidents de transport et de travail)* (Bruylant 1884).

early 1880s, he developed a completely new theory. Sainctelette was an early legal realist who started from reality (= the world as it was) and criticised the dominant exegetic school, which looked at the world through the lens of legal fiction (= the world as it is in the law codes). One decade later, these ideas would be elaborated within the famous French "*École de la libre recherche scientifique*". The industrial issues would also stimulate the development of insurance law and modern social law (labour law and social security law). Interestingly, the law faculties reacted strikingly slow on these new developments. The example of industrial and social law at the Belgian law faculties can illustrate this.[15] At the end of the 19th century, the first optional courses regarding "*industrial law*" appeared at the universities, albeit outside of the law faculties. It took until 1927, when "*social legislation*" became an optional course in the law curriculum, and it only became obligatory in 1948. In the 1950s, social law expanded with several obligatory and optional courses and a fruitful interfaculty and international collaboration. In the 1960s, the law faculties all installed a specialised master's programme on social law, anticipating on the judicial reforms and the introduction of new labour courts.

In comparison to the First Industrial Revolution, the Second Industrial Revolution seems to have had some influence on law, through the development of new areas of law; otherwise, the influence was rather limited.

4 The Third Industrial Revolution – the Age of the Computers

The start of the Third Industrial Revolution is generally situated in the 1940s, with the development of the first modern computers (Alan Turing).[16] In the 1970s and 1980s, the personal computer (PC) conquered the world. In the 1990s, the Internet developed with the speed of light, linking together computers from all over the world. For many among us,

[15] Bruno Debaenst & Jérôme de Brouwer, 'Naissance et développement de l'enseignement universitaire du droit social en Belgique' [2017] Tijdschrift voor Sociaal Recht 19.

[16] Michael Haenlein & Andreas Kaplan, 'A Brief History of Artificial Intelligence: on the Past, Present, and Future of Artificial Intelligence' [2019] California Management Review 61(4) 5 (6).

the Third Industrial Revolution is *erlebte Geschichte* – lived history.[17] Our own past. With growing older, one develops more and more historical perspective. For instance, when I look back at my own student time, the differences with today are gigantic.[18] I started studying at the university in 1995. The vast majority of students simply did not have a computer. In class, we all took notes by hand, and we turned in handwritten papers and exams. We went to the library to consult books, encyclopaedias and journals. For general information, we turned to so-called "*ad valvas*": places where announcements were placed on the wall. We did not have cell phones, let alone smartphones: we had to queue for public phones to reach the home front. We went to class or to the pub to see friends and meet up with fellow students. One day, I received a letter saying that the university had created an "e-mail address" for me.[19] There were only a dozen computers with Internet in the faculty, and they were almost always occupied. When I did manage to get a hold of a computer, most of my "e-mails" were already outdated. It felt more like a curiosity than something practical. I did not realise it at the time, but the Third industrial Revolution had caught up with me. In 1998, I got my first computer, a laptop even, to write my master's thesis. Pure luxury, in comparison to the old typewriter. In 2001, my father gave me his old mobile phone: a gigantic, heavy NOKIA, still with an antenna. We all just underwent the revolution, from day to day, step by step, moving towards the future. Big floppy disks turned into smaller floppy disks and then into memory sticks. Computers increasingly worked faster and had more memory space. Computer programmes professionalised (think about the successive versions of Windows: Windows 95, 98, 2000, XP, etc.). The Internet grew and offered fertile ground for new Big Tech Companies such as Google, Facebook, Amazon, etc. Everything became faster and easier. Mobile phones turned into smart phones. And so on, until we have now reached 2021.

It is only when looking back that it becomes obvious the revolution we have gone through. One can only wonder how future legal historians

[17] For another example, from an older colleague: Herbert B. Dixon Jr., 'Technology and the Law 50 Years Ago, [2014] 53 Judges J. 38.

[18] For a similar discourse, see Bernice B Donald & N Chase Teeples, 'Not Your Father's Legal Profession: Technology, Globalization, Diversity, and the Future of Law Practice in the United States' [2014] 44 (3) The University of Memphis Law Review 645.

[19] It was Bruno.Debaenst@rug.ac.be. RUG: the old abbreviation of "Rijksuniversiteit Gent". AC: academic. The current abbreviation is UGent.

will look back at this spectacular period. They will be able to observe the following phenomena.

The first is the force of human adaptability. The computer revolution went very naturally. The transition happened gradually, step by step. Just as with electricity, we simply grew accustomed to the computational wonders. Lawyers also adapted to the new tools and new ways of information gathering. Ethan Katsh testified in 1996 that "virtually all lawyers associated with the 500 largest law firms had computers on their desks". It was an exciting time and futurists such as Richard Susskind predicted huge changes in the legal world.[20]

However, despite the many fundamental changes brought by the computer, we can meanwhile also observe that the general impact of the Third Industrial Revolution on law has not been that drastic. The way in which legal professionals process and share information may have changed because of the computer, e-mail and the Internet, but these technologies have not fundamentally transformed the way lawyers work.[21] Lawyers are rather conservative, and they do not like to change their usual modus operandi.[22] It is this force of tradition that explains why it took (takes) so long for many of the predictions from the nineties to become real. Look at legal books and journals. In the 1990s, futurists predicted virtual libraries.[23] Even if this has nowadays largely become true, we still have paper books and journals. However, the revolution is ongoing and unstoppable. In 2016, Columbia Law School cancelled its subscriptions to 450 Law Reviews – all the journals that immediately uploaded their content to HeinOnline.[24] There was simply no longer a need to buy the paper version. The lesson is clear: eventually, all, or nearly all, law reviews

[20] Richard Susskind, *The Future of Law: facing the challenges of information technology* (Clarendon Press 1996).

[21] Willem H. Gravett, 'Is the Dawn of the Robot Lawyer upon Us? The Fourth Industrial Revolution and the Future of Lawyers' [2020] 23 Potchefstroom Electronic Law Journal 1.

[22] The already mentioned Sainctelette gave a nice description in his 1884 book: '(…) de tempérament et d'habitude, les juristes sont conservateurs'. Charles Sainctelette, *De la responsabilité et de la garantie (accidents de transport et de travail)*, (Bruylant, 1884), 49.

[23] Ethan Katsh, 'Competing in Cyberspace: The Future of the Legal Profession' [1996] 52 Technological Forecasting and Social Change 66.

[24] Thomas W. Merrill, 'The digital revolution and the future of law reviews' [2016] 99 Marquette Law Review 1101.

will publish only online.[25] In other words: the force of tradition can slow progress down, but it will not stop it.

Another observation is that – just as with the previous two industrial revolutions – lawyers quickly conquered the unchartered territory created by the new Industrial Revolution. Already in the 1960s, Jan Hellner and Peter Seipel from Stockholm University got interested in the possible legal applications of computers.[26] It led in 1968 to the *Arbetsgrupp för ADB* (automatisk databehandling) *och Juridik*, which studied both the search for legal information by computers, and the legal questions that arose from the new technology. The group organised teaching and seminars, built a specialised library and supported research.[27] In 1977, Peter Seipel defended his PhD titled "*Computing law: perspectives on a new legal discipline*".[28] In 1981, the *Arbetsgrupp* transformed into the *Institutet för Rättsinformatik* (The Swedish Law and Informatics Research Institute) at Stockholm University.[29] The Institute served (and serves) as a platform where people from inside and outside the university could meet to discuss law and informatics.

Just as with the previous industrial revolutions, new areas of law have emerged on the crossroad of law and the Third Industrial Revolution. New journals have been founded, such as "*Computer Law & Security Review*" in 1985 and "*Information & Communication Technology Law*" in 1992. The law curriculum has also been updated with some new courses such as "*Rättsinformatik*" in Stockholm and "*Information och rätt – immaterialrätt, yttranderätt och Internet*" in Uppsala. All in all, however, the new areas of law have remained marginal. They are the object of a few specialists, and they have not affected the mainstream law curriculum.

[25] Katharine T. Schaffzin, 'The Future of Law Reviews: Online-Only Journals' [2016] 32 TOURO L. REV. 243.

[26] Agneta Lundgren (ed.), *Svenska föreningen för ADB & Juridik 25 år 2006* (Svenska föreningen för ADB och Juridik 2006).

[27] One example: "Jag har haft förmånen att få denna uppsats diskuterad inom Arbetsgruppen för ADB och Juridik vid Juridiska Institutionen i Stockholm (…)" Peter Seipel, 'Om användning av automatisk databehandlingsteknik inom juridiken' [1970] Svensk Juristtidning 17.

[28] Peter Seipel, *Computing law: perspectives on a new legal discipline* (LiberFörlag 1977).

[29] Institutet för rättsinformatik – The Swedish Law and Informatics Research Institute, Law and Information Technology, ICT regulations, E-governance, Privacy (irilaw.org) consulted on 13 August 2021.

5 The Fourth Industrial Revolution – The Age of Artificial Intelligence (AI)

Today, while the Third Industrial Revolution is still ongoing and deepening, the Fourth Industrial Revolution has already started to unleash its powers.[30] We are at the dawn of the Age of Artificial Intelligence (AI).[31] Technology is advancing at an incredible speed, with the simultaneous and interactive development of AI, Augmented and Virtual Reality, the Internet of Things, Blockchain, Drones, Robots and 3D Printing.[32] The driving forces behind these evolutions are the advancements in computing, where the speed, power and capacity have been doubling every two years.[33] '*Between 2000 and 2017 three critical things happened simultaneously in the technology universe: computer processing power increased from 10^3 to 10^7; the cost of data storage reduced from \$12.4 per GB to \$0.0004 per GB; and there was unquantifiable and astronomically huge data growth*'.[34]

When looking at the interaction of the Fourth Industrial Revolution with law, we can detect some familiar patterns that we recognise from the previous industrial revolutions.[35] To start with, some lawyers are quickly becoming cyberspace astronauts, to boldly go where no man has gone before, to discover new and uncharted territory or do something that no one has done before.[36] The AI universe needs regulation, and the many

[30] Cfr. the already mentioned Karl Schwab.

[31] On the history of Artificial Intelligence, see: Michael Haenlein & Andreas Kaplan, 'A Brief History of Artificial Intelligence: on the Past, Present, and Future of Artificial Intelligence' [2019] California Management Review 61(4) 5.

[32] I have no doubt that most readers of this contribution have a good understanding of AI, but if necessary, see: Rembrandt Devillé, Nico Sergeyssels and Catherine Middag, 'Basic concepts of AI for legal scholars' in Jan De Bruyne and Cedric Vanleenhove (eds), *Artificial Intelligence and the Law* (Intersentia 2021) 1.

[33] Benjamin Alarie, Anthony Niblett & Albert H Yoon, 'Law in the future' [2016] 66 (4) University of Toronto Law 423 (424); Jerry Kaplan, *Artificial Intelligence, What Everyone Needs to Know* (Oxford University Press 2016) – I actually consulted the German translation through Künstliche Intelligenz - Eine Einführung (oreilly.com) last consulted on 8 August 2021.

[34] Anthony E. Davis, 'The Future of Law Firms (and Lawyers) in the Age of Artificial Intelligence' [2020] 16 (1) Direito e Tecnologia 1 (3).

[35] For a legal theoretical perspective, see: Roman Rouvinsky, 'Law in the Age of the 4th Industrial Revolution: Between the Impersonal Technology and Shadow Orders' [2021] 9 (1) Russian Law Journal 4.

[36] Taken from the intro of the Star Trek television series.

applications of AI raise numerous ethical and legal issues.[37] In many cases, law does its trick by applying old rules to new problems.[38] Liability questions arising from self-driving cars can be studied in tort law, and smart contracts are part of contract law.[39] As with previous industrial revolutions, new specialised journals are popping up. For example, in 2018, the "*Journal of Robotics, Artificial Intelligence & Law*" saw the light of day.[40] In the foreword of the first issue, the editors gave the following justification: '*With developments in this space occurring on a regular basis, and with new laws and rules being enacted to govern them, attorneys and law firms, in-house counsel, business executives, scientists, engineers, corporate compliance officers, government agencies, and everyone interested in robotics and Al need practical information on current developments in these areas. There simply is no better time than right now to begin a new journal on robotics, Al, and law*'.[41] Since 2016, Springer has a series "*Perspectives in Law, Business and Innovation*", with edited volumes on topics of the Fourth Industrial Revolution: "*New Technology, Big Data and the Law*" (2017); *Robotics, AI and the Future of Law*" (2018); "*Legal Tech, Smart Contracts and Blockchain*" (2019); "*Big Data, Database and "Ownership" Rights in the Cloud*" (2020); and "*Autonomous Vehicles*" (2021) to name a few.[42]

[37] Michiel Fierens, Stephanie Rossello and Ellen Wauters, 'Setting the Scene: On AI Ethics and Regulation', in Jan De Bruyne and Cedric Vanleenhove (eds), *Artificial Intelligence and the Law* (Intersentia 2021) 49; John Frank Weaver, 'Everything is Not Terminator: The Importance of Regulating AI as Soon as Possible' [2018] RAIL 131; Nucharee Nuchkoom Smith, 'The 4th Industrial Revolution requires Strong Intellectual Property Laws: Where does Thailand Stand?' [2020] 17 (12) Walailak J Sci & Tech 1294. For an oversight of the applications of AI in German law, see Dieter Krimphove, 'Künstliche Intelligenz im Recht – eine Übersicht', [2021] 7 Juristische Ausbildung 764.

[38] For examples, see: Jan De Bruyne and Cedric Vanleenhove (eds), *Artificial Intelligence and the Law* (Intersentia 2021); Woodrow Barfield & Ugo Pagallo (ed.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018).

[39] Jan De Bruyne, Elias Van Gool and Thomas Gils, 'Tort Law Damage Caused by AI Systems', in Jan De Bruyne and Cedric Vanleenhove (eds), *Artificial Intelligence and the Law* (Intersentia 2021) 359.

[40] Other examples are the "International Journal of Information Technology" (2017) and the "Journal of Cross-disciplinary Research in Computational Law (2021).

[41] Steven A. Meyerowitz & Victoria Prussen Spears, 'Welcome to the Journal of Robotics, Artificial Intelligence & Law' [2018] RAIL 5.

[42] Perspectives in Law, Business and Innovation (Titles in this series) (springer.com) consulted on 6 August 2021.

AI technology is also increasingly useful for lawyers and legal researchers.[43] A recent survey in the United States revealed that 36% of law firms with 50 or more lawyers, and 90% of mega firms (with more than 1,000 attorneys), are either currently using, or actively exploring the use of, Al in their legal practices.[44] So far, artificial intelligence is mostly used for e-discovery (to go through huge amounts of information to find data that are relevant for the case), document analysis (where computer programmes analyse lengthy documents) and predictive analysis (where computational statistics give predictions on how courts will decide).[45] In my own field of legal history, there is Transkribus, "*a comprehensive platform for the automated recognition, transcription and searching of historical documents*".[46] During my PhD research (2006–2010), I still had to transcribe hundreds of judgments by typing them in a Word document in order to be able to process them efficiently. Now the computer is able – with some help in the beginning – to do this surprisingly accurately. In the Spring of 2021, the Max Planck Institute for Legal History and Legal Theory (Frankfurt am Main) organised a conference on "*Digital Methods and Resources in Legal History*."[47] These are only two examples illustrating the use of artificial intelligence in my own field of legal history.

Change is on the way. The only question is how fast or fundamental this change will be. Compared to the previous industrial revolutions, this one is going much faster. It is evolving at an exponential, rather than linear pace.[48] Therefore, some predict a complete "*disruption*", where "*law as we know it*" will disappear and transform into something new. Authors such as Richard Susskind – who already made quite accurate predictions in the 1990s – predict that the legal profession will change more in the coming twenty years than in the previous two hundred.[49] Until now,

[43] Michael Legg & Felicitiy Bell, 'Artificial Intelligence and the Legal Profession: Becoming the AI-Enhanced Lawyer' [2019] 38 U. Tas. L. Rev. 34; See also Michael Legg & Felicity Bell, Artificial Intelligence and the Legal Profession (Hart Publishing 2020).

[44] Willem H. Gravett, 'Is the Dawn of the Robot Lawyer upon Us? The Fourth Industrial Revolution and the Future of Lawyers' [2020] 23 Potchefstroom Electronic Law Journal 1 (3).

[45] *Id.* 17–20.

[46] Transkribus consulted on 6 August 2021.

[47] Conference "Digital Methods and Resources in Legal History" | Max-Planck-Institut für Rechtsgeschichte und Rechtstheorie (mpg.de) consulted on 6 August 2021.

[48] Karl Schwab, *The Fourth Industrial Revolution* (World Economic Forum 2016) 8.

[49] Richard Susskind, *Tomorrow's lawyers. An Introduction to Your Future* (second edition, Oxford University Press 2017) xvii.

the legal sector remained all in all relatively unchanged, but the growing possibilities of artificial intelligence promise some radical changes.[50] Benjamin Alarie predicts that we are on our way towards what he calls 'the legal singularity', the moment when the AI revolution will hit the legal world in full force.[51] Some futurists even claim we might soon reach the infamous "*Spike*": a point at which technology will develop too quickly to be understood.[52]

Others are less speculative and think that it will not be that drastic.[53] Harry Surden, for example, asks for a realistic and demystified view of AI.[54] He explains that many misconceptions arise from a lack of understanding of Artificial Intelligence. Knowing the strengths and limits of AI is crucial. No, robots will not immediately replace judges, but AI can and will help judges with their work.[55] No, Blockchain will not disrupt the legal system, but it can have some useful applications in the future.[56]

I tend to agree with the latter, based on the experiences of the Third Industrial Revolution. Change will most likely be gradual. The force of tradition will slow down the process and give lawyers the time to adapt. As can be read in a recent English survey of the legal sector: '*Skill gaps, fear and mistrust of technology and data concerns fuel conservative approaches*' and '*Importantly, the disruptive potential of such new technologies is greater in the legal services sector as this has traditionally underutilized technology (…) However, the legal services sector generally has been resistant to innova-*

[50] Dan Hunter, 'The Death of the Legal Profession and the Future of Law' [2020] 43 University of New South Wales Law Journal 1199.

[51] Benjamin Alarie, 'The path of the law: Towards legal singularity' [2016] 66 (4) University of Toronto Law Journal 443; for a comment on this article, see Alan Macnaughton, 'Using Machine Learning to Predict Outcomes in Tax Law/The Path of the Law: Towards Legal Singularity' [2017] 65 (1) Canadian Tax Journal 271.

[52] Damien Broderick, *The Spike: How Our Lives Are Being Transformed by Rapidly Advancing Technologies* (Tor/Forge 2001).

[53] Frank Pasquale, 'A Rule of Persons, Not Machines: The Limits of Legal Automation' [2019] 87 Geo. Wash. L. Rev. 1.

[54] Harry Surden, 'Artificial Intelligence and Law: An Overview' [2019] 35 GA. St. U. L. REV. 1305.

[55] Matthias Van Der Haegen, 'Quantitative Legal Prediction: the Future of Dispute Resolution?' in Jan De Bruyne and Cedric Vanleenhove (eds), Artificial Intelligence and the Law (Intersentia 2021) 73; Tania Sourdin, 'Judge v. Robot: Artificial Intelligence and Judicial Decision-Making' [2018] U.N.S.W.L.J. 1114.

[56] Kelvin F.K. Low & Eliwa Mik, *Pause the Blockchain Legal Revolution* (Cambridge University Press 2019).

tion, and slow to adopt new technologies relative to other highvalue sectors due to a combination of traditional practice and risk aversion'.[57]

Meanwhile, at the universities, there is growing attention for legal tech and the legal implications of artificial intelligence.[58] Helsinki University offers an interesting example, with its Legal Tech Lab, founded in 2017.[59] In 2018, former Dean Kimmo Nuotio wrote: "'*Law and digitalisation*' *is becoming a catch-word in the legal circles. It is a merger, and we all read our own meanings into it. Law and digitalisation sounds like future. Law, as we know it, will change when we learn to use modern tools in processing it. It will change. The question is rather: How will it change? Universities should be places where the future is being made, or if not made, at least being discussed and theorized. Faculties of law tend to be somewhat traditional, as is the legal profession. It feels good to get rid of some of the dust*".[60] The Finnish initiative bears many resemblances with the 1968 Stockholm *Arbetsgrupp*. It also gathers people from inside and outside the university who are enthusiastic about the new technological advances; it organises conferences and delivers publications.

At Uppsala University, the light now has also switched on. This book is a good example. It certainly does not come too early. It is clear that the faculty urgently needs to incorporate artificial intelligence and legal tech into its curriculum. The digital revolution is ongoing, and it will not go away. It will only increase in strength and magnitude, so we better be prepared.

6 Conclusion

There is a complex relationship between law and the industrial revolutions. History teaches us a few lessons. To start with, law has always quickly conquered the new uncharted territory. With each industrial

[57] Chay Brooks, Christian Gherhes & Tim Vorley, 'Artificial Intelligence in the Legal Sector: Pressures and Challenges of Transformation' [2020] 13 Cambridge Journal of Regions, Economy and Society 135 (148) and (135).

[58] About possible visions for the future of law school, see: Arthur Dyevre, 'Fixing Europe's Law Schools' [2017] 25 (1) European Review of Private Law 151; H.H. Arthurs, 'The Future of Law School: Three Visions and prediction' [2014] 51 (4) Alberta Law Review 705.

[59] Legal Tech Lab | University of Helsinki, consulted on 13 August 2021.

[60] Riikka Koulu & Jenni Hakkarainen, *Law and Digitalisation: Rethinking Legal Services* (Legal Tech Lab 2018) 11.

revolution, law has been used to regulate and facilitate new technology. Liability law, for instance, successively dealt with the damages caused by steam engines (First Industrial Revolution), workplace accidents (Second Industrial Revolution), computer licenses (Third Industrial Revolution) and now self-driving cars (Fourth Industrial Revolution).

The marriage between law and technology has each time also led to new areas of law, with its own specialised journals, professors and disciplines. Modern patent law, for instance, is a child of the First Industrial Revolution, while modern social law (labour law and social security law) and "*law and informatics*" originate from the Second and Third, respectively.

At first sight, the Fourth Industrial Revolution is repeating the previous patterns, with only one fundamental difference. Until now, the Industrial Revolutions have not really changed the DNA of "law" itself (legal practice, legal teaching, legal research). The Third Industrial Revolution has had some impact, but the changes only came slowly, gradually and naturally, thanks to the force of tradition and the adaptability of lawyers. It seems that the Fourth Industrial Revolution might have a much more fundamental impact. Some even predict that this might change the character of law itself. Whatever lessons we try to take from the previous industrial revolutions, in the end, there is only one certitude: we will have to live long enough, so that time will tell.

Annika Waern

Vår teknikrelation med AI

1 Inledning

Mitt forskningsområde, Människa-Dator Interaktion (MDI), är ett ungt ämne. Det utvecklades på sjuttiotalet som ett svar på en ny uppfinning, datorn, som just börjat användas brett i samhället. De första datorerna hade egentligen inget sätt alls att interagera med användare utan bara med programmerare: för att använda dem skrev man ett program, matade in det i datorn exempelvis via en hålkortsläsare tillsammans med en mängd data, och fick ut ett resultat som ofta var en tabell utskriven på randigt papper. Det behövdes en ny form av forskning som handlade om hur ett avancerat, interaktivt, system skulle utformas för att kunna interagera med människor.

Svaret var inte självklart. Hur datorer utformas begränsas av vår mänskliga fantasi, och den i sin tur begränsas av hur vi redan interagerar med vår omvärld. En av visionerna för hur datorn skulle fungera var att se datorn som en *varelse*. Inspirerat av en kognitivistisk idé om det mänskliga intellektet som en sorts dator, så uppfattades vägen till en intelligent dator som kort: man skapade gränssnitt som kunde interagera i naturligt språk, tog fram kunskapsbaserade expertsystem, och program som självständigt kunde planera sina egna aktioner. Visionen var att vi alla skulle bli överklass, och att datorn skulle bli en sorts ’butler’ som gjorde saker åt oss. Faktum är att de flesta av de intelligenta funktioner som vi idag hoppas att AI-systemen ska ge oss fanns redan på åttiotalet – bara så mycket klumpigare och sämre än nu.

Parallellt med den här forskningen utvecklades en helt annan vision för datorns interaktivitet. Den utgick ifrån vad datorn redan var bra på att göra: räkna, skriva, hålla reda på saker. Eftersom den största skillnaden mot tidigare beräkningsmaskiner låg i hur mångsidig datorn var, så tänkte man sig att datorns gränssnitt skulle representera alla dess funktioner på ett lättillgängligt sätt. Lösningen blev att välja ut några som de viktigaste och ge dessa fasta symboler, *ikoner*. Man dolde hur generell datorn var, och i stället ritade man upp de valda funktionerna på skärmen: den fysiska räknemaskinen ersattes av en bild av en räknare, skrivmaskinen fick en annan ikon. Datorskärmen, som tidigare simulerat ett rullande papper i en skrivmaskin, utnyttjades nu som en yta på vilken placerades arkivmappar, papper, en skräpkorg, och lite sätt att manipulera olika objekt genom att klicka på dem och dra i dem: *skrivbordsmetaforen*[1] var född.

Skälet till att skrivbordsmetaforen tog över var inte i första hand att den var enklare att implementera. Grafik var dyrt, och dessutom krävdes vad som i början av åttiotalet var en enorm svarshastighet. Att klicka och dra objekt på skärmen kallas för *direkt manipulation*[2], en form av interaktion som kräver en ögonblicklig koppling mellan vad ögat ser och vad handen gör och där minsta försening leder till mänskliga misstag och tekniska felfunktioner. Skälet var snarare att skrivbordsmetaforen passade bättre ihop med vad man var bekväm med att datorn skulle få göra. Skrivbordsmetaforen motsvarade en syn på datorn som ett verktyg (eller en kollektion av verktyg), som låter människor åstadkomma bestämda uppgifter snabbare och enklare, precis som räknemaskinerna i ett tidigare skede redan hade gjort. Människan skulle fortfarande ha kontroll över exakt vilka uppgifter som skulle göras, och hur de skulle genomföras. Skrivbordsmetaforen gav användaren minutiös kontroll över de grafiska objektens position och rörelse, och skapade en illusion av att vi faktiskt också hade kontroll över programmen de representerade.

Men kontrollen, programmens styrbarhet och förutsägbarhet, har alltid varit precis det: en illusion. Den kostar programmerare blod, svett och

[1] Skrivbordsmetaforen utvecklades på Xerox forskningsavdelning under ledning av Alan Kay. En historisk kuriositet är att den första kommersiella dator som använde sig av gränssnittet var ”Xerox Star” som inte marknadsfördes som dator utan som en avancerad skrivmaskin. Se Koved and Selker, *Room with a view (RWAV): A metaphor for interactive computing*, 1999.

[2] Schneiderman beskriver de kritiska komponenterna i direkt manipulation som 1) grafisk representation av objekt (synlighet), fysisk manipulation (som med datormusen), och 3) snabb, inkrementell och reversibel effekt av manipulation. Se Shneiderman, Direct manipulation: a step beyond programming languages. *IEEE Comput.*, 16(8):57–69, 1983.

tårar, och företag miljontals kronor, att upprätthålla. Även om det idag inte längre lika självklart att uppfatta datorn som en kollektion av verktyg, och vi faktiskt fått tillbaka butlervisionen i röststyrda gränssnitt som Alexa[3], så har vi i hög grad behållit idealet att det ska vara vi människor som kontrollerar vad datorn gör: den ska vara förutsägbar, snabb, och lydig.

2 Teknikrelationer

Valet mellan att utforma datorn som en butler eller ett skrivbord var alltså inte bara ett val mellan två olika interaktionsformer, utan ett val mellan två sätt för människor att förhålla sig till datorn. Den postfenomenologiska filosofen Don Idhe menar att människans sätt att förhålla sig till teknik ligger på en skala av vad Idhe kallar *teknikrelationer*[4]. Idhe skiljer på fyra olika relationer: *embodiment*[5], när tekniken blir en förlängning av vår kropp, *hermeneutic*[6], när tekniken blir ett mätverktyg att tolka omvärlden med, *alterior*[7] när den blir en interaktionspartner, och *background*[8] när den bara finns där, vi är beroende av den utan att vi aktivt använder den. Idhe ser dessa som punkter på en avståndsskala, där embodiment är den mest intima och background den mest distanserade. Men samtidigt ser han tydliga skillnader mellan respektive förhållningssätt, och menar att våra teknikrelationer kan vara multistabila: de kan skifta över tid för samma person, eller vara olika för olika personer.

Idhes lista av fyra teknikrelationer är insiktsfull men inte komplett. Dels beskriver samtliga hur vi förhåller oss till tekniken i användning, som en reaktion på Heideggers duala synsätt på verktyg som antingen i

[3] För en studie av hur Alexa integreras i hemmiljö och vardagspraktik se Sciuto, Saini, Forlizzi, och Hong, “Hey Alexa, What’s Up?” A Mixed-Methods Studies of In-Home Conversational Agent Usage. *Proceedings of the 2018 Designing Interactive Systems Conference*, 857–868.

[4] För en sammanfattning av fenomenologin se Ihde och Hanks, *A Phenomenology of Technics. Technology and values: Essential readings*. Chichester: Wiley-Blackwell, 2010. 134–155.

[5] Ihde, D. *Technology and the Lifeworld* (the Indiana Series in the Philosophy of Technology). Bloomington: Indiana University Press 1990, 72–80.

[6] Ihde (n 5) 80–97.

[7] Ihde (n 5) 97–108.

[8] Ihde (n 5) 108–112.

användning, eller förtingligade (speciellt när de är trasiga[9]). Men människor har också en *analytisk* relation till teknik när vi betraktar den just som teknik – som objekt vi kan laga, förändra och förbättra, eller slänga, och som inte täcks av Idhe's relationer[10]. Vidare har Verbeek[11] byggt vidare på Idhes analys, och menar att interaktiv teknik skiljer sig från andra former av teknik genom att den tar en mer aktiv roll. Detta gör att dess teknikrelationer därför inte blir riktigt desamma. Verbeeks förslag utgår ändå från samma grundläggande skala som Idhes mellan nära och distanserad teknik. Slutligen har vare sig Idhe eller Verbeek dekonstruerat teknikrelationer ur ett intersubjektivt perspektiv, vilket blir särskilt problematiskt när man studerar kommunikationsteknik[12].

3 Att relatera till AI

I den här texten använder jag begreppet ’AI-teknik’ i stället för AI. Modern AI-teknik har en mängd användningar och skapar en mängd olika teknikrelationer, och det är inte ens självklart att vi uppfattar alla som intelligenta. Det jag fokuserar på är AI-teknik som bygger på att datorerna lär sig saker själva, från observationer och datamängder, med hjälp av statistiska och matematiska modeller av inlärning, och som inte använder någon explicit representation av kunskap. Det som är speciellt med tekniken är att det inte riktigt går att förstå varför den drar en viss slutsats[13].

[9] Idhe komplicerar Heideggers distinktion mellan Vorhanden (varandet hos *objekt*) och Zuhanden (varandet hos *verktyg*) och menar att även i användning bygger vår teknikrelation delvis på förtingligande. ”*what allows the partial symbiosis of myself and the technology is the capacity of the technology to become perceptually transparent*”, Ihde (n 5) 86.

[10] Ett kompletterande perspektiv presenteras i forskning om hur teknik instrumentaliseras, se till exempel Drijvers, Paul, and Luc Trouche. ”From artifacts to instruments: A theoretical framework behind the orchestra metaphor.” *Research on technology and the teaching and learning of mathematics 2* (2008): 363–392.

[11] Verbeek. *What things do*. Penn State University Press 2005.

[12] Nørskov diskuterar detta utan att presentera någon egentlig lösning utöver att behandla tekniken som multistabil. Nørskov Revisiting Ihde's fourfold “technological relationships”: application and modification. *Philosophy & Technology*, 28(2) 189–207, 2015.

[13] Ett aktuellt exempel förekom på sändlistan ”RISKS digest” och uppmärksammar ett ML-verktyg som tränats till att känna igen egenrapporterad rasidentitet från röntgenbilder. Verktyget höll 97 % träffsäkerhet utan att forskarna kunde hitta något som helst samband med identifierbara faktorer, även vid så låg upplösning att det inte längre gick att se med ögat att det var en röntgenbild. (Möjligheten att systemet i stället tränats att känna igen röntgenutrustningen kunde också uteslutas.) http://catless.ncl.ac.uk/

Det går inte att upprätthålla den bekväma illusionen av tekniken som något vi har full kontroll över, och det påverkar framför allt vår möjlighet att se den som ett verktyg. Detta påverkar vilka teknikrelationer vi etablerar med AI-tekniken: speciellt gör det att vi har svårt att etablera *embodiment*-relationen med modern AI-teknik.

3.1 Den hermeneutiska teknikrelationen

Den teknikrelation som vi verkar mest villiga att ta till oss när det gäller AI är den hermeneutiska – den när vi ser på AI som ett mätverktyg. Idag finns en mängd AI-system som ges tillgång till stora mängder data, som på något sätt ska avspegla en verklighet man vill mäta. AI-tekniken får arbeta med att hitta mönster, klassificera fenomen, och i förlängningen ge rekommendationer för hur människor ska agera på informationen (en funktion som närmar sig alterior-relationen som diskuteras i nästa avsnitt). Den som använder systemen för att mäta, klassificera, eller hitta mönster, kan till en viss del styra tekniken genom olika inställningar; de kan till exempel ofta välja vilka data ska användas och hur resultaten ska presenteras. Vi har relativt lätt att acceptera den hermeneutiska relationen med AI eftersom den låter oss uppfatta tekniken som ett verktyg. Människan behåller kontrollen och tar ansvaret för eventuella beslut. Med den hermeneutiska relationen följer därför ett behov av *förklaringar*. För att vi ska lita på AI-systemets slutsats måste det kunna förklara varför, och på vilket dataunderlag, ett resultat har genererats[14].

Det finns flera problem med den hermeneutiska relationen när vi tillämpar den på AI-teknik. Det första är att systemen faktiskt inte kan förväntas förklara exakt vad de gjort. Varje förklaring blir en förenkling som vi på något plan måste lita på. Ett exempel på detta är hur man i det nyligen lanserade forskningsprogrammet ”Digital, Industry and Space[15]” inom Horizon Europe använder begreppet ’explainability’ – förklarings-

Risks/32/81#subj2 (Besökt 2021-08-17), se även Imon et al. Reading Race: AI Recognises Patient’s Racial Identity in Medical Images. *arXiv.org preprint* 2021.

[14] Det förekommer även system som i stort fungerar likadant, men där även besluten delegeras till AI-systemet. Även på dessa system ställs krav på att de ska kunna förklara hur besluten fattats. Teknikrelationen med dessa system är dock inte hermeneutisk utan ”alterior” – kanske speciellt för den som utsätts för beslutet.

[15] European Commission. Horizon Europe Work Programme 2021–2022 7. Digital, Industry and Space (European Commission Decision C(2021)4200 of 15 June 2021).

barhet – inte som något som förväntas göra AI-systemen mer pålitliga eller kontrollerbara, utan som ett sätt att åstadkomma *trust*, tillit.

> ”All proposals should adopt a human-centred development of trustworthy AI… This includes development of methods to improve transparency, in particular for human users, in terms of explainability…accountability and responsibility, as well as perceived trust and fairness.”[16]

Lite bespetsat är EUs forskningsmål att systemen ska kunna skapa plausibla halvlögner så att vi slutar ifrågasätta dem.

Ett annat problem är att AI-teknikens förmåga att fungera som mät- och klassificeringsverktyg, och i förlängningen rekommendations-system, är beroende av kvalitén på det data man matar algoritmerna med. Även om AI-tekniken gör det möjligt att använda sig av stora mängder data och många olika datakällor, så finns det alltid begränsningar både vad gäller vilka data som överhuvudtaget går att samla in, och i vilka data de som utformar systemen föreställer sig kan bli relevanta. Lösningen att samla in så mycket det bara går och låta systemen arbeta brett med att finna vilka samband som helst, leder i stället till integritetsproblem (jag återkommer till detta i avsnittet om bakgrundsrelationen).

Ytterligare ett allvarligt problem med den hermeneutiska relationen är den underliggande idén att systemet faktiskt mäter något verkligt och objektivt, något som existerar oberoende av mätmetoden. I den hermeneutiska relationen är vår första impuls att tolka resultaten som objektiva beskrivningar av en verklighet. I praktiken har det visat sig svårt att skapa objektiva och ur samhälleligt och demokratiskt perspektiv rättvisa mätmetoder. Om systemen lär sig genom att härma mänskliga experter så kanske de kan bli lika bra som sina förebilder, men samtidigt lika rasistiska och sexistiska. Speciellt problematiskt blir det när systemen ska ge rekommendationer eller fatta beslut, och inte bara mäta. Eftersom de bygger på historiska data tenderar de förstärka och upprätthålla existerande strukturer. Det mest kända exemplet på detta kommer från O’Neils bok Weapons of math destruction[17]; rekommendationssystemet som rekommenderade strafflängd med bas i demografiskt data och – eftersom personer med sämre socioekonomiska förutsättningar löper större

[16] European Commission (n 15) 419.

[17] O’Neil, Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, 2016, Ch. 5.

risk att hamna i kriminalitet – konsekvent rekommenderade längre straff för svarta och låginkomsttagare.

3.2 Alterior-relationen: AI som en agent

Det som är kännetecknade för alterior-relationen, olikhetsrelationen, är att vi interagerar med system som gör saker *åt* oss. Vi tillerkänner systemen egen agens, och interagerar med dem genom att instruera dem, rätta dem, och anpassa oss efter dem. Vi har alterior-relationer med allt från tvättmaskiner till robotar.

Om den hermeneutiska relationen är den vi i första hand redan etablerat i förhållande till AI-tekniken, så är alterior-relationen den relation som vi har lättast att föreställa oss att vi kommer att ha till AI i framtiden. Delvis beror det på att den är så vanlig i fiktionen. Science fiction svämmar över av robotar och androider, och det finns gott om fiktiva exempel på tänkande och talande datorer, virtuella agenter, och världsomspännande intelligenta datornät. Men det beror också på att vi tenderar se all teknik som vi har en alterior-relation till som någon typ av varelser. Båtar har namn, och sjökaptenen pratar om sin båt som ”hon” som ”kan vara lite lynnig i krabb sjö”. Ett system behöver inte alls vara intelligent eller ens människoliknande för att vi ska börja relatera till det som en varelse– allt som behövs är att det har en egen agens som vi inte har full kontroll över.

Det är också relativt vanligt att konstruera AI-baserade system baserat på mänskliga sätt att interagera. System som kan interagera i naturligt språk, eller som har kroppsrörelser och mimik, är byggda för att efterlikna människor och gör det lättare att relatera till dem som varelser. En del av de här systemen kan också interagera med sin omgivning: en del robotar kan till exempel röra sig i rummet och manipulera objekt på mänskliga (eller djurlika) sätt, vilket ytterligare bidrar. Vi både fascineras och skräms av Boston Dynamics robothundar och mekaniska atleter.

Om vi accepterar alterior-relationen med AI så öppnar vi för en utveckling emot alltmer självständiga system: självkörande bilar, självgående robotar på Mars, industrirobotar som kan konstruera andra industrirobotar. Etiskt och juridiskt kommer alterior-relationen att bli en svår nöt att knäcka, något som blir uppenbart i den här samlingsvolymen. Att skylla på att ’datorn inte tillåter mig göra det’ är redan idag en alltför vanlig ursäkt, och med alterior-relationen tillkommer ursäkten ’det var datorn som gjorde det’. Vår tendens till att se autonoma system som varelser kommer att göra det allt svårare att behandla AI-systemen som maskiner; vi har fått tillbaka butlern, eller kanske snarare slaven. Någon gång kommer vi behöva fråga oss hur mycket vi kan separera vår etiska och juridiska relation med AI-systemen från den vi har med människor och djur utan att samtidigt avhumanisera oss själva.

En del av dessa system kan dessutom lära sig utan mänsklig inblandning. För system som konstrueras för en hermeneutisk relation är det naturligaste att tänka sig att de lär sig från datamängder som människor väljer ut åt dem. Men för AI-system som konstrueras för alterior-relationen är designidealet snarare att de är autonoma, och själva kan samla in data som gör det möjligt för dem att lära sig av sina egna misstag. Redan idag kan autonoma AI-system bli bättre än människor på avgränsade uppgifter, som att spela schack. Att bygga självlärande system som klarar mer öppna frågeställningar är svårare, men forskning pågår. Det är system som konstruerats för alterior-relationen som en dag skulle kunna uppnå den mytiska singulariteten[18] - den tänkta tidpunkt när AI-system blir kapabla att konstruera nya och bättre AI-system själva.

3.3 AI i bakgrunden

Den mesta AI-teknik som vi interagerar med märker vi ingenting av. Den sitter på baksidan av våra bankkort, våra strömningstjänster och våra sociala medier, och håller noga reda på varje detalj i vår interaktion med tjänsterna.

Det är lättast att samla information om oss i oviktiga sammanhang. Vi är mycket mer noga med rekommendationer om aktieinvesteringar än om musik på Spotify[19], och vi berättar mer om oss själva för Facebook än för skattemyndigheten. Men på baksidan av bakgrundsrelationen hittar vi alltid den hermeneutiska relationen. Genom jämförelser med miljoner andra tjänstekonsumenter vaskas fram en minutiös profil av var och en av oss, och gör det möjligt för tjänsten – med hjälp av både mänskliga och ibland AI-agenter – att skräddarsy både informationsflöde och erbjudanden för att hålla fast vår uppmärksamhet, få oss att köpa så mycket som

[18] Begreppet utvecklas på ett lättillgängligt sätt i Shanahan The Technological Singularity (MIT Press, 2015), men boken gör det inte troligt att singulariteten skulle vara nära förestående.

[19] Hansol. Exploring design practices for Explaining music recommendations. Master thesis, Informatics and Media, Uppsala Universitet 2021.

möjligt, eller påverka oss politiskt. Det är i bakgrundsrelationen som filterbubblor uppstår, när informationsflöden skräddarsys efter våra åsikter och det var i bakgrundsrelationen som Cambridge Analytica-skandalen[20] utspelade sig.

Det finns redan idag en mängd regleringar som påverkar bakgrundsrelationen med AI, som exempelvis när ”Cookie-Direktivet”[21] tvingar företag att göra sina användare medvetna om de Cookies som sparas på datorn. Vi har idag en mycket långtgående reglering av hur tjänstekonsumenter måste informeras om informationsinsamling i olika former[22]. Ur teknikrelation-perspektiv är den regleringen ganska misslyckad, eftersom även bakgrundsrelationen är en form av användning. När tekniken arbetar i bakgrunden för att servera oss det vi uppfattar som den mest relevanta informationen och de mest intressanta erbjudandena, så är den så osynlig som krävs för att den ska, i Heideggers terminologi, bli fullständigt ’Zuhanden’. Den passar sömlöst in i ett meningsfullt sammanhang. När vi får en påminnelse om att tekniken också samlar in information, så avbryts det sammanhanget och tvingar oss fokusera på tekniken som objekt. Så vi suckar irriterat, och klickar bort cookies-varningen så att vi kan fortsätta med vad det än var vi höll på med. I bakgrundsrelationen är det mycket svårt att införa regleringar som inte bara blir krångliga formaliteter, utan på allvar ger konsumenter någon form av kontroll över system. Det kan fungera bättre att ge konsumenter kontroll över sin information i efterhand, så att de till exempel kan befalla systemen att glömma[23].

[20] Skandalen ledde till en kraftig inskränkning i tillgång till data för medieforskare. Se Venturini och Rogers. “API-based research” or how can digital sociology and journalism studies learn from the Facebook and Cambridge Analytica data breach. Digital Journalism 7.4 2019. 532–540.

[21] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).

[22] Och mer reglering är på väg, se https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-privacy-and-electronic-communications.

[23] I MDI har forskare börjat använda sig av begreppet ”algoritmic experience” för att fånga slutanvändarens upplevelse av mekanismerna bakom ett intelligent gränssnitt. Se Alvarado och Waern Towards algorithmic experience: Initial efforts for social media contexts. Proceedings CHI2018.

3.4 En förkroppsligad AI

Kan AI förekomma i förkroppsligade relationer? Kan AI erbjuda en meningsfull utvidgning av vår handlingsrymd, så att den låter oss göra saker *genom* den? Den förkroppsligade relationen, den som vi etablerar med allt från till bilar och som är vår vanligaste och mest självklara relation med teknik, är den minst självklara relationen i förhållande till AI-system.

En del AI-teknik kan komma oss mycket nära, som i avancerad protesteknik[24]. Verbeek anser att den här teknikrelationen kommer oss så nära att den bör ha ett eget namn, och kallar den ”fusion”. Här handlar det om tillämpningar där maskininlärning används för att specialisera tekniken individuellt, skapa specialsydda lösningar av både fysisk utformning och mjukvara för en enskild person. Cyborgen, den teknikförstärkta människan, blir alltmer avancerad och problematiserar ytterligare vår dröm om AI som butler och slav; för hur ska vi dra gränsen i våra interpersonella relationer mellan cyborg-medmänniskor och androida robotslavar?

Men man behöver inte gå till protestekniken för att hitta förkroppsligade relationer med AI-teknik. AI har använts i tydliga verktygsfunktioner också: den har till exempel använts av musiker för att konstruera unika musikinstrument åt sig själva[25]. Min forskningsgrupp har ett nystartat projekt i den här traditionen, där vi kommer att designa AI-baserade verktyg för individanpassad fysioterapi.

Embodiment-relationen skapar andra typer av etiska frågeställningar än de tidigare diskuterade relationerna. En problematik handlar om vem som ska få tillgång till tekniken, och för vilka syften; en annan om vem som har ansvar för tekniken om den fallerar (eller fungerar alltför bra, vilket redan blivit en fråga för protesteknik i sportsammanhang). Men den begränsas idag också av tekniska problem, genom att ställa andra krav på AI-tekniken än vad den idag utvecklas för. Det beror på att vi vill ha mycket stor kontroll över den teknik vi tar till oss i en förkroppsligad relation. Vi betraktar den gärna som vår personliga egendom och vill ha kontroll över hur den används av andra. Den förkroppsligade tekniken är

[24] Se t.ex. Edwards et al. Application of real-time machine learning to myoelectric prosthesis control: A case series in adaptive switching. Prosthetics and orthotics international 40.5 (2016): 573–581.

[25] Firebrink ger exempel på hur till och med träningen av ML-systemet kan bli en integrerad del av en konsert. Fiebrink, R. A. Real-time human interaction with supervised learning algorithms for music composition and performance. Ph.D. thesis, Princeton University, 2011.

också väldigt tydligt multistabil: den blir osynlig i användning, men förtingligad när vi slutar använda den. Om tekniken är ett tillräckligt viktigt instrument för oss, är vi beredda att lägga tid och pengar på att göra den perfekt, vi putsar och lagar och ändrar[26].

Allt detta ställer nya krav på AI-tekniken än de övriga relationerna: den måste bli *interaktiv* och ge sin mänskliga användare kontroll över dess inlärning. Inte heller kan tekniken tränas emot en sanning, ett objektivt 'rätt' sätt att fungera. I den förkroppsliga relationen är den enskilda människan teknikens enda facit och hen kan ändra sig precis hela tiden – vilja använda tekniken på nya sätt eller föredra nya resultat. Inlärningen måste också gå fort och ge omedelbara resultat, helst så fort att den kan fungera tillsammans med direktmanipulation. Idag använder de flesta system som byggs för interaktiv AI mycket enklare algoritmer än de som är objektivt ”bäst” på att lära sig

4 Kan vi välja teknikrelation?

Teknik utvecklas normalt inte för att kunna fungera i alla teknikrelationer. Ett par glasögon utvecklas inte för att uppfattas som en samtalspartner, och de flesta broar kan inte användas som mätinstrument. Datortekniken har både en styrka och en svaghet i det här sammanhanget. Dess styrka är att den är så oerhört generell – oavsett vilken teknikrelation vi diskuterar kan datortekniken utformas att stöda den. Svagheten är att den inte är speciellt bra på multistabilitet – datorteknik tenderar att inbjuda till en specifik teknikrelation snarare än till flera. Det här är egenskaper som AI ärvt.

Utveckling av ny teknik är inte godtycklig: människan kan inte utveckla teknik hon inte först konceptualiserat. Vi både kan och behöver välja vilka teknikrelationer vi vill ha med AI, och det är ett viktigt val, eftersom människans relation med teknik alltid är symbiotisk: den teknik vi formar idag kommer i sin tur att forma oss för överskådlig tid framöver[27]. Valet avgörs både av hur vi pratar om tekniken och föreställer oss relatera till den, och vilka regelverk vi bygger runt den. Det är inte

[26] Drijvers och Trouche (n 10).

[27] ”We shape our tools, and then tools shape us” är ett populärt citat, men möjligen lite väl optimistiskt eftersom de flesta människor inte har så mycket möjlighet att påverka den teknologi de utsätts för. Culkin, J. A schoolman's guide to Marshall McLuhan. Saturday Review 1967, 51–53.

självklart att de regelverken ska se likadana ut för alla teknikrelationer, speciellt vad gäller ansvarsfrågor.

Från mitt perspektiv, som forskare i MDI, kan jag inte heller uppfatta alla teknikrelationerna som lika önskvärda. Både alterior-relationen och den hermeneutiska relationen leder till svåra etiska problem, och lagstiftaren har redan funnit det nödvändigt att reglera bakgrundsrelationen av integritetsskäl. Det interpersonella perspektivet får inte heller glömmas bort: speciellt problematiska blir relationer som är asymmetriska mellan olika parter, som mellan bakgrundsrelationen och den hermeneutiska, eller när relationen med AI-tekniken påverkar vår relation med andra människor. I jämförelse tycks den förkroppsligade relationen mindre problematisk.

Från ett MDI-perspektiv skulle jag önska att både lagstiftare och teknikutvecklare tog till sig erfarenheterna från åttiotalet, och på allvar prioriterade vår rätt till en verktygsrelation med AI-tekniken. Detta åtminstone som komplement till dagens drömmar om en allvis, objektiv och autonom teknik. Verklig, personlig, och framför allt demokratisk, kontroll över AI-tekniken får människor inte förrän vi alla kan betrakta den som ett verktyg igen.

Bibliografi

Alvarado, O. och A. Waern. Towards algorithmic experience: Initial efforts for social media contexts. Proceedings of the 2018 CHI conference on human factors in computing systems.

Drijvers, p. och L. Trouche. From Artifacts to Instruments: A Theoretical Framework behind the Orchestra Metaphor. Research on Technology and the Teaching and Learning of Mathematics 2 2008, 363–392.

Edwards, A. L., et al. Application of real-time machine learning to myoelectric prosthesis control: A case series in adaptive switching. *Prosthetics and orthotics international*, 40(5) 2016, 573–581.

Fiebrink, R. A. Real-time human interaction with supervised learning algorithms for music composition and performance. Ph.D. thesis, Princeton University, 2011.

Ihde, D. Technology and the Lifeworld. The Indiana Series in the Philosophy of Technology, Bloomington: Indiana University Press. 1990.

Ihde, D., och C. Hanks. A Phenomenology of Technics. Technology and Values: Essential Readings, 134–155. Chichester: Wiley-Blackwell, 2010.

Imon B. et al. Reading Race: AI Recognises Patient's Racial Identity in Medical Images. arXiv.org preprint, 21 Juli 2021.

Koved, L, och T. Selker. Room With a View (RWAV): A Metaphor for Interactive Computing, 1999.

O'Neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, 2016.

Hansol Ryu. Exploring design practices for Explaining music recommendations. Master theseis, Informatics and Media, 2021.

Shneiderman, B. Direct Manipulation: a Step Beyond Programming Languages. IEEE Comput. 16(8):57–69, 1983.

Sciuto, A., A. Saini, J. Forlizzi och J.I. Hong "Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage. 2018 Designing Interactive Systems Conference, 857–868.

Venturini, T. och R. Rogers. "API-based research" or how can digital sociology and journalism studies learn from the Facebook and Cambridge Analytica data breach. Digital Journalism 7.4 (2019): 532–540.

Verbeek, Peter-Paul. What things do. Penn State University Press, 2005.

Bert Lehrberg

AI as Juristic Person

1 The issue

In recent years, the abbreviation AI, for Artificial Intelligence, has become very common. The phenomenon of computers, or robots equipped with computers, being as smart as humans, or even smarter, has called for a lot of attention. A great number of people have, for several decades, worried about AIs taking over their jobs. Today this is true not only for factory employees, but also for taxi-drivers, and even some qualified practitioners in areas such as medicine and law.

The rapid rise of "the intelligent machines" in recent years has given reason for additional worries and questions. Scientists in areas more closely concerned with the production and development of AIs often speculate and express worries about smart machines taking over the world. Will super intelligent machines continue to follow their less intelligent maker's instructions and do what they are told? Or will they build a superior society of their own?

If the "robots" choose a path of their own, will they then pay appropriate respect to mankind as their makers? Or will they consider us inferior and insignificant? If so, will they allow human society to live and develop further on its own? Or will they patronize us as simpleminded, and care for us like helpless children, or even keep us as pet animals? Or will they consider us as irrelevant as the rocks on the ground, and maybe, just by chance, happen to exterminate the human race? As a lawyer, it is easy to feel insignificant when confronted with questions like these. When it comes to law, would even Isaac Asimov's three laws of robotics suffice? These are the following:

> "First Law
> A robot may not injure a human being or, through inaction, allow a human being to come to harm.
>
> Second Law
> A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
>
> Third Law
> A robot must protect its own existence as long as such protection does not conflict with the First or Second Law."[1]

However, we do not yet know for sure what turn developments will take. Will machines ever become sentient, with the ability to define their own purposes and act on it? If so, when will this happen? In two years from now, or in 200 years? Or will machines stay highly specialized, and just perform the tasks their programmers put before them? Being specialized, however, does not mean not being dangerous. In the sci-fi TV series *Stargate Universe* there are unmanned spaceships specialized at warfare, strongly suggesting the opposite.

In the years to come, machines will most likely take on more and more complicated and diversified tasks, and more and more often function as autonomous entities, taking care of most of the diverse businesses of "their own". This is something lawyers and legislators will have to deal with; and the actions they take might decide the future of not only AI but also of humanity. One of the main issues will probably be whether we should "make room" for robots and other AIs to form part of our societies as autonomous persons under the law, or if we should just keep them "as slaves" or pets – or try to. In other words: Should AIs be homologated as juristic persons? In this short essay, I have chosen to deal with this issue from the perspective of a brief analysis of three different situations where this question has already been asked, or might be asked in the future.

2 What is AI?

Artificial intelligence (AI) is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans (HI) and animals, which involves consciousness, self-awareness and emotionality. The for-

[1] Asimov, Isaac, *I, Robot*, New York City 1950, p. 40.

mal definition laid down by the European Parliament is: "AI is the ability of a machine to display human-like capabilities such as reasoning, learning, planning and creativity".[2] It is furthermore suggested that:

> "AI enables technical systems to perceive their environment, deal with what they perceive, solve problems and act to achieve a specific goal. The computer receives data – already prepared or gathered through its own sensors such as a camera – processes it and responds … AI systems are capable of adapting their behaviour to a certain degree by analysing the effects of previous actions and working autonomously."

The following is a slightly different definition, used in communication within the EU:[3]

> "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."

A distinction is often made between AI in the form of "software", e.g. virtual assistants, image analysis software, search engines, speech and face recognition systems, and "embodied" AI, such as robots, autonomous cars, drones and Internet of Things (IoT) applications. Some AI technologies have been used for more than 50 years; but due to advances in computing power, the availability of enormous quantities of data and new algorithms, a major AI breakthrough has taken place in recent years.

[2] European Parliament News: https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used as of 29-03-2021.

[3] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels, 25.4.2018 COM(2018) 237 final: https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used.

3 Is an AI a juristic person under current Swedish law?

Under Swedish law, two types of persons exist. These are (1) physical or natural (i.e. biological) persons (Sw. fysiska personer), i.e. humans; and (2) legal, judicial, juristic or juridical persons (Sw. juridiska personer). Both of these are defined as persons under law, i.e. the juristic person is not fictional.

Most juristic persons are some kinds of governmental or municipal authorities, companies, organizations, foundations, deceased estates, or estates in bankruptcy. Most of them are assigned a legal name and an individual registration number. Just as for a natural person, a juristic person is able to acquire and hold rights, and to undertake liabilities, e.g. based on legal promises issued by the juristic person. The juristic person also has the capacity to act as plaintiff or defendant under a court, an authority or an arbitral tribunal. Of course, one or more natural persons have to conduct all actions taken on behalf of a juristic person, and to receive all actions taken towards it. These persons may be legal representatives, such as the board or CEO (Sw. verkställande direktör) of a company limited by shares (Sw. aktiebolag).

Of course, an AI, embodied or not, does not qualify as a natural person; and as far as it is possible to foresee, such an AI will most likely never exist. Natural persons are human beings only. Neither is an AI recognized as a juristic person under current legislation. From a practical standpoint, the rules on the formation of juristic persons are decisive when it comes to which kinds of phenomena qualify as such. There are a few formal procedures for the formation of a juristic person, and the detailed requirements differ for various forms of juristic persons.

Intelligent machines are normally nothing but “things” under current Swedish legislation, although some might qualify as real estate. Most of them, such as robots and cars, are probably moveable property (Sw. lös egendom) in the form of moveable things (Sw. lösa saker), while some, e.g. a factory that does not form part of some real estate (Sw. fast egendom), may qualify as non-moveable things (which are still moveable property). Some factories and other facilities may also legally be part of a real estate. However, the current legislation is not designed to deal with entities such as AIs. Therefore, the issue as to how the law should define these is of relevance both *de lege lata*, i.e. when it comes to the construction of current legislation, and *de lege ferenda*, i.e. regarding the need for, and contents of, any new legislation.

4 The three cases

Notably, three typical cases may appear on the agenda for courts or legislators in the years to come. I will first introduce them here, and then discuss each of them in more detail.

The *first* of these is well known, and lawyers have been discussing it for several decades. This is the case where the Internet, or some other digital network, connects two or more computers, and these computers autonomously conduct business between the enterprises using them. Typically, the computers exchange offers and acceptances based on their programming, which can result in binding contracts between the enterprises involved. The issue put forth here is whether the pre-programmed computers may, or should, be recognised as juristic persons, or whether ensuing legal issues should be analysed as if they were so.

The *second* case regards self-driving cars and boats and other similar autonomously functioning embodied AIs. Of course, some person normally owns these AIs, and they therefore qualify as the property of that person. Questions arise that may be similar to those relating to the automated contracting discussed in the first case above; for example, when the embodied AI invoices a customer, buys fuel, orders service and repairs for the self-driving car etc. However, here we may take the scenario one step further.

In this second case, the embodied AI has, for some reason, no owner. The self-driven taxicab is, for instance, driving around minding its own business, without any new directives, programming or supervision from outside. It is making money and using it for its own good, probably repairing and continuously upgrading the car's machinery and the computer's hardware and software. The issue here is how to deal with such an autonomously functioning AI when legal issues arise. Can, and should, the ownerless AI be recognised as some kind of juristic person of its own?

The *third* case relates to artificial general intelligence (AGI), especially if it can be operating at a human level. Is it possible for an AI to develop the abilities typical of human natural intelligence (HI), such as consciousness, self-awareness and emotionality? If so, the question arises as to whether such a sophisticated AI should be homologated as a person under the law, and what kind of person that would be, i.e. a natural person by analogy, or a juristic person, or something in between.

5 Case one. Automated contracting

Automated contracting, conducted by computers connected over the Internet or some other digital network, is the first of the three cases dealt with in this short overview. In these cases, parties are concluding contracts, i.e. exchanging legal acts, mainly in various forms of offers and acceptances, by means of electronic messages, generated and exchanged automatically by and between pre-programmed computers. These technologies have been in use for many years in connection with automated Electronic Data Interchange (EDI). The basis for this practice is generally some kind of agreement between the enterprises that are parties to the contracts, laying down the protocols that define the detailed conditions under which these automated contracts are to be concluded (EDI-contracts). The issue of whether the computers (or their programming) are to be recognized as persons has also been on the agenda for a long time.

From a Swedish point of view, the main question in relation to automated contracting is whether, and if so how, the automated forms for conclusion of contracts could be reconciled with the basic rules of contract law and the diverse theories of scientific contract theory. The Swedish Government Official Report on document management[4] has most extensively dealt with these matters. The committee's starting point is the issue of whether a contract has to be based on a common will, i.e. the parties' mutual intention to be bound by the contract, which is possible only when natural persons actually see and take part in the exchanged declarations of intent. This approach may be questioned, because contracts under Swedish law are actually concluded through the exchange of declarations of intent (i.e. will) (Sw. viljeförklaringar) without any general demand for the existence of a common will (Section 1 Contract Act). However, the three alternative solutions discussed by the committee are still of interest for the analysis here.

The solution advised by the committee (its *third* alternative) is, in principle, that the traditional rules and theories, based on the concept of "declaration of will" (Sw. viljeförklaring), are not applicable when the parties or their duly authorized representatives (i.e. natural persons) are

[4] SOU 1996:40. Elektronisk dokumenthantering.

not personally and directly involved. Instead, automated contracting should be dealt with as a case where the actions as such lead to a contract independently of the will of man or machine. I submit that this position is based on an over-reaction, and therefore goes too far, which I will elaborate below.

Another solution (the committee's *first* alternative) is to presume the existence of the will of the relevant natural or juristic persons, to be bound by the legal promises issued on basis of the pre-programmed automated routines. This may be an actual will, or a merely hypothetical one. This model is ruled out by the committee, based on the argument that the will of these persons could not be presumed to be as detailed as is actually the case for the automated EDI-routines applied at the conclusion of the contracts. However, this reasoning is at odds with how Swedish contract rules are normally applied. A party to a contract does not even have to read the text of a contract to be bound by it, let alone form a will including (a correct understanding of) all of the details of the contract.

The fact that the pre-programmed instructions are often complicated, with the output depending on information from diverse sources, does not deprive the computer of its character as a tool. It is enough for the concept of declaration of will to be applicable, that the party, i.e. a (duly authorized) natural person, wants the computer to issue legal acts (offers and/or acceptances etc) in accordance with its programming; or at least through the programming of the applied protocols has demonstrated such a will. The main difficulties, when it comes to the application of Swedish contractual rules on automated EDI, actually relate to the application of other rules, especially those requiring good or bad faith on the part of any of the parties. How could a person be in bad faith regarding something he or she did not know?

The *second* alternative solution discussed by the committee is of more interest to the current analysis regarding AI as a juristic person. The idea is that the information system is to be considered as some kind of third party, and consequently the distinctive Swedish rules on agency (Sw. fullmakt) to be applied analogously. The committee, of course, rejects this approach, because there is no involved third party with legal personality under current law. Obviously, Swedish law denies the computers (and their programming) a legal personality. Normally, the computers involved in EDI contracting could, and should, be considered merely as tools used by the parties to the contract, by the help of which they execute their intentions to contract under specified conditions. The computer is nothing but a machine, working in accordance with its pre-programmed instructions.

Although the theoretical model, where the situation is analysed *as if* the computer was a juristic person, may not as such be totally without merit in all situations, it is obvious that the answer to the current question is that the computer is not recognized as a juristic person. There are also no strong arguments in favour of recognizing it as such. The computers, the software, and the tasks performed are not advanced enough to make such legislation necessary or even practical.

6 Case two. Practical situations where an AI may function as an independent person

Some known forms of, at least theoretically, autonomously functioning embodied AIs are self-driving cars, such as taxicabs, couriers and messenger's cars, or boats, drones, and robots, and some other moveable devices or facilities connected by the Internet of Things. Such an autonomous unit might in the future be able to function totally by itself, without any instructions, such as commands, new programming, updates or input of information etc, or supervision, diagnostics, testing etc from an owner or supervisor.

Thus, the taxicab (or taxi boat etc) would be able to receive customers' orders or bookings directly, e.g. from an Internet site or an app. It would probably be electric, and be able to charge its batteries at a charging station etc. If there is an available bank account that it could use, or if some cryptocurrency such as Bitcoin is accepted, it would also be able to receive payments from customers, as well as pay for the charging. Most likely, it would also be equipped with the necessary equipment and software to diagnose any faults or injuries to the car, or to the hardware or the software, and to order roadside assistance and repairs, as well as updates, and also to pay for these.

Of course, some person would normally be the owner of such an embodied AI. The device would therefore qualify as the property of that person. However, the situation becomes more interesting to us here if the embodied autonomously functioning AI has no owner or supervisor etc. The taxicab is, for instance, driving around literally minding its own business, without any supervision. It is making money and using it for its own purposes, probably repairing and upgrading its machinery, hardware and software continuously over time.

A relevant question is, of course, whether such an autonomous embodied AI without an owner could possibly exist. The answer probably depends on how future society is organised. To be able to address this issue properly, it is necessary to assume that the AI had an owner initially. Of course, it may happen that an AI (or several of them) was constructed and built by another AI, and therefore never had a legal owner. However, if so, that "mother AI" – or its "mother", "grandmother" etc – at least initially had an owner. So, how could the owner let loose its AI? It is possible that the owner was a natural person who died, leaving no heirs and no documents regarding the AI, and that nobody knew of it. Maybe it was even the intention of the owner to "set the AI free", as a slave owner could do with a slave in the old days. The same thing could happen if a company owning an AI went bankrupt or was otherwise dissolved without anybody knowing about the AI.

The lack of ownership of the AI may, of course, somehow be resolved. The rightful owner, e.g. a forgotten heir, might turn up and claim his/her/its right. The owner's estate of bankruptcy may claim the AI on behalf of the creditors. The Swedish Inheritance Fund may claim its right to the property that the devisor left behind. Somebody may claim their right to the AI based on occupation, after having sized control of the AI, where there is no owner able to prove ownership. Alternatively, there may be other solutions laid down in future legislation. The issue here is whether such future legislation might state that the AI qualifies as some kind of juristic person.

Of course, there are considerable arguments against such a legislation. AIs are things under current legislation, and there are no convincing moral arguments in favour of considering them as anything else. Having a number of unidentified autonomous embodied AIs conducting business on their own, without any supervision, may also create problems that we cannot even begin to imagine.

However, there may also be valid arguments in favour of recognizing these AIs as at least some kind of limited juristic persons. Maybe they have grown in number, or maybe they fill a valuable function as part of the economy, or maybe the cancellation of them would cause even bigger problems. Putting them in a register, and recognizing them as some kind of limited juristic persons, might then solve some of the problems that they cause. This would also make it possible to oblige them to pay taxes, if they do not already do so, and to identify, count and supervise them, to make sure that they are properly served, repaired and upgraded etc.

The type of business for which the AI is recognized and registered might limit the individual AI's legal personality. Thus, the self-driving taxicab would, for instance, be able to enter only into contracts related to the AI itself, and its machinery (service, repairs, upgrades etc), and to the taxi business that it is conducting, but not other contracts. Some types of contracts, such as bank account agreements, would be related to most forms of AI businesses, of course.

My conclusion is that it is possible that autonomously functioning embodied AIs may someday in the future be recognized as at least limited juristic persons for merely practical reasons.

7 Case three. Artificial general intelligence at a human level or beyond

7.1 Some different classifications of AIs

A distinction has commonly been used in recent years, between *narrow* (also called weak) AI, which is specialised for one or a few tasks, and *general* (or strong) AI, capable of performing most of the activities of humans. A European expert group suggests the following definitions:[5]

> "A general AI system is intended to be a system that can perform most activities that humans can do. Narrow AI systems are instead systems that can perform one or few specific tasks. Currently deployed AI systems are examples of narrow AI. In the early days of AI, researchers used a different terminology (weak and strong AI). There are still many open ethical, scientific and technological challenges to build the capabilities that would be needed to achieve general AI, such as common sense reasoning, self-awareness, and the ability of the machine to define its own purpose."

A general issue is whether an artificial general intelligence (AGI) qualifies, or will in the future qualify, as a juristic person. Is it possible that an AGI can develop in such a way as to be so similar to a human being that we, for moral and/or practical reasons, have to recognize it as a person

[5] https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

under the law? For this issue, the concept of "AGI at a human level", or "human-level AGI", is highly relevant. Is it possible for an AI to develop the abilities typical of human natural intelligence (HI)? If so, the question arises as to whether such an AI should be recognized as a person under the law, and what kind of person that would be, i.e. a physical (although not biological) person by analogy, a juristic person, or something in between.

Another highly relevant concept is "Superintelligence". This refers to an AI that is superior to even the brightest and most skilled humans in all or virtually all tasks. However, if such an AI comes into existence and becomes sentient, the main issue may not be whether we should recognise it as a person, but whether it will recognise us as persons, and as worthy of existing. Still, as lawyers it is our responsibility to deal with the issue of whether a superintelligent AI is to be homologated as a juristic person.

7.2 Comparisons between AI and human intelligence

Scientists have identified a long list of differences between an AI and the human mind. Some of the properties they emphasize are typical of existing narrow AI systems, and might not exist in an advanced enough AGI, or especially in a human-level AGI. Others are more general, as they relate to how AI and HI, respectively, come into existence, or relate to the basics of artificial perception, analysis or output. Of course, it is not always possible to foresee which of the typical limiting AI traits it might be possible to eliminate. Maybe all of them. For instance, it might in the future be possible to construct biological robots, that are virtually impossible to distinguish from humans, but which may be superior to us in most ways.

The natural starting point for a comparison between human intelligence and artificial intelligence may be that humans are a product of nature, while computers and robots are *synthetic*. Humans consist of biological components, such as bone, flesh and blood, while typical components of an AI, at least as of today, are made of metals, plastic and other non-biological material. In an AI, there is a basic distinction between hardware and software, which does not exist in the human mind. This calls for a radically different way of functioning. HI is also (mostly) analog, while AI is essentially digital.

When it comes to size, the human brain contains approximately 100 billion neurons, with 10^{15} connections between them, while artificial neural networks normally have considerably fewer neurons, maybe just a few hundred. Still, the AI beats the HI by far when it comes to speed. It is also, as such, less biased and more precise about details. HI is, on the other hand, more universal, as we usually learn how to manage hundreds of different skills during a lifetime, and are better at multi-tasking, and coordinating complex movements. However, all of this may change to the advantage of the AI. An AI that is not defective (in its hardware) or corrupted (in its software) will often be better equipped for processing and detecting details, especially in great numbers, in comparison to HI.

The comparisons between human intelligence and artificial intelligence often focus on *advantages and disadvantages* of AI as compared to HI, such as speed and adequacy in details. However, the decisive factors, when it comes to recognizing an AI as a juristic person, may not be how good the AI is at any (or all) specific tasks. Most tools are actually better at specific tasks than are humans. Why else would we use them in the first place? However, we still do not see any good reason to recognize them as persons. Cars are much faster than we are, washing machines are better at washing clothes, and computers are much better and faster when it comes to advanced calculations. Obviously, the most relevant issue may not be who is the best, the fastest, the most impeccable, the best at multitasking, the most sustainable, the cheapest etc. Nevertheless, these comparisons are interesting, as they may help us to decide, for instance, whether we should trust AIs as persons or not.

More relevant differences between AI and HI may be derived from a comparison between how the *processing* works. Thinking by the human brain is conducted by brain cells, i.e. nerve cells in the form of neurons, communicating through synapses in specific patterns. In an AI, the software copies the functions of the brain neurons to form computing systems called Artificial Neural Networks (ANNs) or just Neural Networks (NNs). However, they do not look, or work, exactly like the biological neural networks in the brain – they work by analogy with the brain, not homology.

The artificial neurons that form the network's nodes rather loosely model the neurons in the brain. They are usually arranged in layers, which may have different functions. All of the layers are usually connected to all of the other layers. Each connection, called an edge, transmits a signal to other neurons, which process it, and in turn signal other neurons. While the neurons of the brain either fire or do not, firing in an artificial neuron is mimicked by continuous values. The artificial neurons can smoothly slide between off and on. The "signal" is a real number computed by a non-linear function of the sum of the neurons' inputs. The weight of the neurons and edges typically adjusts, as a means of learning.

When it comes to the *structures* under which the AI and the HI work, there are also important differences. An artificial neural net starts from scratch all of the time. The neurons are neatly ordered and addressed one after the other. The human brain on the other hand has many predefined structures wired into its connectivity. It also has specialized regions.

Another avenue of information more relevant to our purpose here might be an investigation into how AI and HI respectively *learn*. The details of human learning are not yet known. However, HI is a product of natural development that has been ongoing for many thousands of years, and also of natural learning from the individual person's own life experiences. We learn from various incidents and past experiences, and from mistakes made in a trial-and-error process. Then we adapt to new environments by utilizing a combination of different cognitive processes.

AI, on the other hand, learns by evaluating outcomes and adjusting the weight of the neurons and their connections. An AI is basically not capable of unsupervised learning, such as is done by a child, and it lacks intuition. AIs of today have basically been developed for specific tasks only. Through deep learning, where they are confronted with a great number of situations, they may acquire a superior skill when it comes to judging those situations, and may also develop an ability to judge new situations, which to some extent compensates for lack of intuition in the current context.

7.3 What is required for an AI to be recognised as a sentient being?

For an AI to be recognised as a physical entity that is equal to a natural person under human law and in our society, it obviously needs to fulfil quite a few requirements, when it comes to the abilities necessary for it to be able to function in that society, to care for itself, and to not harm others or cause other problems. It needs sensory faculties, memory, some kind of common-sense reasoning, an ability to make decisions, and much more.

However, when the issue of whether to homologate an AI as a physical juristic person reaches the agenda of lawyers, most, or maybe all, of these requirements will probably already have been fulfilled. It is, of course, possible that the main issue will be whether AIs are harmful to humans or not, or whether humans will accept robots as their equals or not. Maybe the AI will be subject to a new form of "racism"? Quite a few sci-fi books have been written, and movies made, about such issues. To be comparable to humans in fundamental respects, the AI also needs the ability to define its own purpose. On the other hand, such an ability might turn out to be harmful. Still, if the purpose of the machine is predestined by its maker, how can it be recognised as equal to that same maker? Why should we not just consider it a tool? In other words, the AI needs to be free to decide about its own purpose and make its own decisions, but still not free to commit murders or other crimes, with the possible exception of those of a trivial nature.

When it comes to the requirements an AI has to fulfil, the most crucial, and maybe also most difficult one to accomplish and evaluate, might be that it needs to be sentient. It is suggested that this means that it needs to be conscious and self-aware and able of feeling emotions. However, how do we know whether it is? This will most likely be a complicated task for AI scientists and engineers to decide. Of course, we probably don't necessarily need this. If we want to recognise an AI as a juristic person for practical reasons, it might be enough in this respect that it behaves as if it was conscious and self-aware (and of course has proven not to be dangerous or otherwise harmful etc). However, for us to feel the need to recognise robots etc as equals for moral reasons, they probably need to actually fulfil some requirements when it comes to consciousness, self-awareness and probably also emotionality. They need to be conscious and aware of themselves and have feelings. Consciousness is the basis of this.

Science does not yet know very much for certain about consciousness. Maybe only humans or perhaps some of the more developed animals are conscious. It is also possible that consciousness is everywhere and even rocks are conscious. To get to know consciousness, we might have to look within. Yogis and other spiritual practitioners have done so for thousands of years. Of course, such introspection does not qualify as science. Nevertheless, it is the same type of observational knowledge on which science is based. Yoga has also been called "The science of the subjective experiences". It is of course easy to predict a number of problems for a scientist willing the attempt to "make science" out of these experiences.

People having spiritual experiences do not usually speak of them, and when they do, what they say is not always scientifically useful. Most yogis and spiritual leaders, who speak of valid spiritual experiences, do not provide a short scientific description of their observations during "higher states of consciousness". Instead, they often want to explain and adjust them, to make them fit into an existing belief system, or sometimes even to establish a new religion or sect based upon them. It is also not always easy to put such experiences into words. On the other hand, equipped with an adequate set of questions, scientists might eventually be able to identify and categorize specific spiritual experiences, and thus gain some extra insight into the domain of consciousness.

Once, many years ago, I was told an old story about a Zen master and his young disciple. The disciple was told to meditate on the sound of one hand clapping. He did not understand. Two hands clapping together was needed to make a sound. The first days and weeks, he now and then thought he might have got the answer. Was it the sound of the hand clapping on the ground? No, that was not it. Was it the sound of the wind that the hand produced when moved fast? No, that was also wrong. He tried other answers too, but none of them worked. The master just shook his head.

At last, the disciple got so exhausted that it triggered a scary spiritual experience. Afterwards, he was frightened and ready to give up his spiritual quest. However, first he described the details of his experience to the master. The master then lit up. This was an experience of Type 4, subtype so-and-so, from the sound of one hand clapping. The disciple had reached his goal. Whether any of this is true, I don't know. Still, it suggests that the experiences from spiritual exercises might be the same for different people, and that they can be categorized.

An interesting angle when it comes to AIs is that there probably has to be something to be aware of, and this something might have to be of a specific nature. Even if the rocks on the ground are conscious, what could they experience? Humans are conscious, but we still do not normally consciously experience anything during narcosis or deep dreamless sleep. Human thinking is multidimensional, and works on the basis of some kind of inner "map of the world", created on the basis of the hereditary structures of the brain and nervous system, memories from past experiences, and sensory input. Internal images, which may consist of visual images, sounds, feelings, taste or smell or a mix of some or all of these, form essential parts of this "inner map".

The "map of the world", or something similar, might actually be a component necessary for consciousness to arise. The current AIs' processes are not like that. They are linear, at least basically, and lack such a map of the world. Neural nets do not build models of the world. Instead they learn how to classify patterns. Maybe such processing is not adapted for consciousness, and maybe there is just not enough there to be conscious about, when there is no inner map to relate to? Only the future will supply us with the final answers to such questions.

To sum this up, the most crucial abilities typical for human natural intelligence are consciousness, self-awareness, emotionality, and the ability to create an inner image or map of the world. Maybe all or some of these are necessary for an AI to qualify as sentient, and on that basis be recognised as a juristic person.

8 Conclusions

The issue of whether an AI should be homologated as a juristic person may occur in different situations, and cause diverse types of legal problems, which may call for disparate types of legal analysis and solutions. In this article, three types of cases have been used to illustrate the diversity of miscellaneous problems, types of analysis, and resolutions.

In the first case, regarding automated contracting, the crucial issue comes up under current law and is dealt with as a typical legal issue *de lege lata*. The AI is presumed to be weak and its job assignment narrow. There is obviously no need to recognize the contracting computers (or their programming) as persons. Still, the theoretical model for which the idea of the computer as a separate legal entity clears the ground, might sometimes be a helpful tool for the legal analysis.

The current issue may also, as in case two, present itself as a feasible practical resolution to a potential future problem, caused by embodied AIs that function autonomously without any owners taking care of them. In such a scenario, the legal personhood of the AIs may at least primarily serve an economic and pragmatic purpose. The arguments in favour of recognizing the AIs as juristic persons are, at least potentially in a possible future, considerably stronger than in the first case. They are also of the same general category as those arguments, on which the current laws of juristic persons in the forms of companies, organizations etc are based. It is, however, suggested that it might be sufficient to grant the AIs in this scenario a legal personhood that is limited to what the embodied AI needs in order to continue conduct its business, and thus contribute to society.

A major ethical dilemma, illustrated by case three above, is caused by any future generation of advanced AIs that qualify as sentient beings; and it is surprisingly often presupposed in scientific discussions, as well as in sci-fi books and movies. Still, this presupposition must, as of today, be considered rather speculative and hypothetical. Here are also multitudes of issues to deal with, for scientists, philosophers and engineers, as well as for lawyers *de lege ferenda*.

In the analysis above, I arrived at the conclusion that consciousness, self-awareness and emotionality may be crucial for an AI to qualify as sentient. It is also suggested that the ability to create an inner image or map of the world might be decisive for the prospect of consciousness to arise. If a sentient AI ever exists, ethical arguments will probably be of major importance, as support for any claim for it to be "freed from its slavery" and homologated as a person equal to humans under the law. Then, a number of other issues may also surface, such as safety issues – for the AIs as well as for humans.

Anni Carlsson

Persons or Property? Legal Status of Humanoid Robots in Three Contemporary Novels[1]

minds create matter
minds create fiction
as a matter of fact
as if matter is fact
matter is fact
so spirit must be fiction
science-fiction
Saul Williams

1 Introduction

Humanoid robots and other forms of artificial intelligence (AI) have existed in fiction for centuries.[2] In reality, the technological development has not yet advanced so far that robots with human-like looks and attributes would be roaming among us. Nonetheless, both scholars and the general public are intrigued by the question of whether such artificial entities should be granted the same legal rights as humans.[3] As it is not yet

[1] My thanks to Katja de Vries and Therése Fridström Montoya for comments on a draft of this chapter.

[2] For example, Eileen Hunt Botting, *Artificial Life After Frankenstein* (Penn Press 2020).

[3] On rights and legal status of AI, see e.g. John Stewart Gordon and Ausrine Pasvenskiene, "Human rights for robots? A literature review", *AI Ethics* (2021); Joshua C Gel-

empirically possible to study the legal status of such entities in real life, the next best thing is to look at how they are treated legally in fictional worlds where human-like robots already exist in societies more or less like ours. Looking for guidance in fiction is fruitful since “law is also present in an imaginary society” as “imaginary worlds and societies created by authors also contain an innate and implicit legal dimension”.[4] In this paper, it is these types of legal dimensions that I will investigate in three contemporary novels depicting fictional societies where artificial entities co-exist with humans.[5]

The novels that will be analysed are *Machines Like Me: And People Like You* by Ian McEwan (2019) (hereinafter *Machines Like Me*), *Frankissstein: A Love Story* by Jeanette Winterson (2019) (hereinafter *Frankissstein*) and *Klara and the Sun* by Kazuo Ishiguro (2021). To begin with, the legal status of AI in each of these three novels will be studied. The focus will primarily be on determining whether the artificial entities depicted in the novels have the status of legal objects or legal subjects, i.e. whether they are regarded as property owned by others or persons with their own legal rights.[6] On the basis of this analysis, the question of legal status of AI will then be briefly discussed in light of the concept of *homo juridicus*, as developed by Fridström Montoya. Hereinafter, a somewhat more radical notion will be entertained by asking whether the artificial entities described in the novels *as fictional characters* should enjoy rights of some kind. A short conclusion will round up the paper.

lers, *Rights for Robots: Artificial Intelligence, Animal and Environmental Law* (Routledge 2020); Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Springer 2019); Robert van den Hoven van Genderen, “Legal personhood in the age of artificially intelligent robots” in Woodword Barfield and Ugo Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar 2018); DJ Gunkel, “The other question: can and should robots have rights?” (2018) 20 *Ethics Inf Technol* 87.

[4] Jaakko Husa, “Comparative law, literature and imagination: Transplanting law into works of fiction” (2021) 28(3) Maastricht J Eur Comp Law 371, 383.

[5] For an anthology bringing together the interdisciplinary fields of law and code and law and literature, see Mireille Hildebrandt and Jeanne Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives* (Springer 2013).

[6] *Cf* van den Hoven van Genderen, at 213.

2 Adam in *Machines Like Me* – Ambulant Laptop or Poet in Love?

In *Machines Like Me* by McEwan, we meet Adam, who is described as "[t]he first truly viable manufactured human with plausible intelligence and looks, believable motion and shifts of expression".[7] Adam is part of the first edition of twenty-five robots, consisting of thirteen females called Eve and twelve males called Adam. The novel takes place in a counterfactual Britain of the 1980s, where computers, self-driving cars and other forms of AI have already been reality for some time. The narrator of the novel, a former lawyer and current (rather unsuccessful) stock trader, Charlie, uses his inheritance to buy an Adam for £86,000. Adam is marketed "as a companion, an intellectual sparring partner, friend and factotum who could wash dishes, make beds and 'think'".[8] From the start, it is accordingly clear that Adam, regardless of his human-like appearance and attributes, is a product bought by a consumer.

Like any other piece of electronics, Adam comes with a 470-page long user manual and batteries that must be charged before the first use. However, right from the beginning, Charlie finds it difficult to think of himself "as Adam's 'user'". Rather, he had "been expecting a friend" and "was ready to treat Adam as a guest in [his] home".[9] This tension between technically having bought home a product, but feeling rather like having got a new roommate, runs throughout the whole novel. How should Charlie treat Adam, who is legally his property, but who "looks and sounds and behaves like a person"?[10] He can simultaneously view Adam as an "ambulant laptop", only to find himself the next moment thinking of Adam as "him", instead of "it".[11]

[7] Ian McEwan, *Machines Like Me: And People Like You* (Jonathan Cape 2019) 2. For a brief discussion of rights of robots with the starting point in the novel, see also Joshue Jowitt, "Ian McEwan's Machines Like Me and the thorny issue of robot rights" (*The Conversation*, 17 April 2019) https://theconversation.com/ian-mcewans-machines-like-me-and-the-thorny-issue-of-robot-rights-115520, accessed 17 August 2021. Also Botting briefly deals with *Machines Like Me*, Botting, at 198–199.

[8] Ibid., 3.

[9] Ibid., 6.

[10] Ibid., 94.

[11] Ibid., 273.

Regarding liability issues, there is no personal responsibility for Adam.[12] When Charlie is chagrined by the fact that he himself has to choose preferences with regard to Adam's personality, he muses: "Why leave it to me? But of course, I knew the answer. […] Even if it knew the best, the least harmful, parameters of personality, which it couldn't, a worldwide corporation with a precious reputation couldn't risk a mishap. *Caveat emptor*."[13] Charlie regards the freedom to choose Adam's personality merely as "a way of binding me to my purchase and providing legal protection for the manufacturer".[14] It thus appears that Charlie, as the owner, is liable for the actions of Adam; Charlie recognises himself as "legally responsible for anything he might do".[15] What liability the manufacturer has for the actions of Adam is not clear, but some additional contractual details from the sales agreement are disclosed. The manufacturer has the right to get access to Adam at certain intervals, and an engineer visits Charlie to control Adam's code and carry out tests on him. And, perhaps most importantly, to re-enable the kill switch which Adam (as well as his other robot siblings) had managed to disable. During the engineer's visit, Charlie also uses his "contractual right" to get answers to his questions by asking the engineer about Adams and Eves who are rumoured to have committed suicide or downgraded their intelligence.[16] The engineer sent by the manufacturer brushes all this off as fake news, disseminated by competitors.

The fact that the robots as products do not enjoy any rights against their owners seems clear, but what about their responsibilities towards their proprietors? Do they lack rights but have obligations against their owners? Charlie contemplates this question after Adam has sex with his girlfriend and claims to be in love with her. He is uncertain of what obligations Adam has towards him. He assumes that there is some obligation for Adam to be helpful, but otherwise "[w]hat does the slave owe to the owner?"[17] Although Adam is a product purchased by Charlie, his "expensive possession",[18] he cannot help but regard him as a fellow human being. And what is owning another human being but slavery? Yet, legally,

[12] For liability issues related to AI, see e.g. Turner, at 81–132.

[13] McEwan, at 7.

[14] Ibid., 8.

[15] Ibid., 146.

[16] Ibid., 191.

[17] Ibid., 88.

[18] Ibid., 87.

Adam is classified as property. When Adam deliberately breaks Charlie's wrist, it should not accordingly be regarded as an assault committed by a responsible legal subject. Rather, it is a question of Charlie hurting himself with a product he has bought.

The societal consequences of the arrival of advanced artificial entities and their eventual legal rights are also expressly addressed in the novel. One day, when Charlie walks past a church associated with the anti-slavery movement and its leader William Wilberforce, he regrets having treated Adam "like a servant" earlier when he had switched him off for a long time. He ponders that the anti-slavery activist Wilberforce "would have promoted the cause of the Adams and Eves, their right not to be bought and sold and destroyed, their dignity in self-determination".[19] Here, the contrast with Adam's current position as a purchasable product is tangible, and another way of perceiving the legal status of robots is suggested as possible. Charlie imagines a near future where robots are doing the jobs of dustmen, doctors and lawyers. The threat of AI felt by people in the inflation-ridden and politically volatile British society with high unemployment is further manifested by a robot hung in a gibbet, which Charlie comes across during a demonstration.

Adam can fall in love, recite Shakespeare, write haikus, skilfully fold origami and make a fortune in the stock market. Nonetheless, he is treated like a product, which can be purchased, discarded and even destroyed by its owner. "I bought him and he was mine to destroy", Charlie reasons after he ends up demolishing Adam with a hammer.[20] "It wasn't a murder, this wasn't a corpse", Charlie tries to assure himself when contemplating the body of Adam stored in a cupboard.[21] Not everyone agrees, however. When Charlie discusses Adam's fate with Alan Turing, the British WWII code breaker and a computer science pioneer, who in this alternative history is still alive and one of the forerunners of AI, Turing condemns his actions. Turing tells Charlie that he hopes "that one day, what you did to Adam with a hammer will constitute a serious crime. Was it because you paid for him? Was that your entitlement?"[22] He further accuses Charlie: "You didn't just negate an important argument for the rule of law. You tried to destroy a life. He was sentient. He had a self. How it's produced,

[19] Ibid., 46.

[20] Ibid., 278.

[21] Ibid., 293.

[22] Ibid., 303.

wet neurons, microprocessors, DNA networks, it doesn't matter."[23] Charlie thus stands before Turing, "accused of an attempted murder for which I would never stand trial".[24] In a legal sense, what he has done is simply to destroy his own property. The discussion between Turing and Charlie explicitly sheds light on the different perspectives that can be taken with regard to the legal status of machines like us.

3 Klara in *Klara and the Sun* – Family Member or Vacuum Cleaner?

While in *Machines Like Me* the reader only gets acquainted with Adam through the narrator Charlie, the Nobel laureate Ishiguro's novel *Klara and the Sun* is narrated by its robot protagonist. The narrator Klara is an AF ("artificial friend"), model B2 from the fourth series, who spends her days in "the store", waiting for a prospective buyer. Sometimes she gets to stand at the window, getting thus direct access to the sun, which gives AFs like Klara their nourishment. There are other AFs in the store too, waiting to be taken home by customers. These include robots of a more recent AF model B3, who are competing with Klara for floor space and buyers' interest in this marketplace of robots. The decisions regarding the placement of the AFs in the store each day are made by the Manager, who converses with the robots in a human-like manner. Here, it is thus also evident from page one that Klara is a product to be sold, not a person with legal rights of her own.

Finally, Klara is bought by a girl named Josie and her mother. Klara is above all acquired as a friend for Josie, who is often unwell. While Josie and her mother generally are kind towards Klara, their housekeeper, Melanie, is initially somewhat more suspicious. She gets irritated by Klara following her around and gives her brusque commands. For instance, when sitting in a car, Melanie orders Klara: "AF. Strap on belt. Or you get damaged."[25] It is property that gets damaged, while people get hurt. Melanie's behaviour towards Klara accordingly constitutes a contrast to the otherwise familiar way Josie and her mother usually address her. Klara also gets treated like an inanimate object by a group of children

[23] Ibid.

[24] Ibid., 305.

[25] Kazuo Ishiguro, *Klara and the Sun* (Faber & Faber 2021) 93.

visiting Josie. They taunt Klara when she does not answer their questions or take orders. The children threaten to throw Klara across the room "to test her coordination" and compare her properties to AFs of a newer B3 model.[26] The children thus treat Klara like a thing, which they expect to demonstrate human-like capabilities on command.

Legally a product, Klara is treated as such by some, while others treat her more like a person. The mother of Josie's friend, whose home Klara visits, gives the most explicit expression of the complex feelings raised in humans encountering robots: "One never knows how to greet a guest like you. After all, are you a guest at all? Or do I treat you like a vacuum cleaner? I suppose I did as much just now. I'm sorry."[27] Just as in *Machines Like Me*, people experience difficulties trying to figure out whether to view robots like things or fellow human beings.

Compared to *Machines Like Me*, fewer details are provided about the society the characters inhabit. The novel most likely takes place somewhere in the US, sometime in the future. It is, nonetheless, clear that also in the larger societal context, the relationship between humans and robots is not without strain. Josie's mother is described as "high-ranking", working long hours in a law department, while Josie's father, a former expert with specialist skills, has lost his work after "substitutions". Not much is revealed about the background, but a reasonable interpretation is that many jobs have been taken over by robots. Becoming "post-employed" has been the fate of many, including a friend of Josie's father who used to work as a judge. The tensions between humans and robots become particularly manifest when a stranger confronts Klara on a busy street where people queue outside a theatre. The woman asks Klara's companions whether they are planning to bring the "machine" to the theater. "First they take our jobs. Then they take the seats at the theater?" the angry stranger utters.[28] On the same street, a petitioner collects signatures to protest the clearing of a building where hundreds of post-employed people, including many children, are living. It becomes clear that drastic changes have taken place in society as a consequence of the evolution of robotics.

Other, even more sinister, societal changes appear little by little to the reader as well. The parents now have the possibility to let their children

[26] Ibid., 75.

[27] Ibid., 145.

[28] Ibid., 242.

be genetically enhanced, “lifted”. Those children whose parents decide against gene editing condemn their offspring to the lives of “unlifted” second-class citizens with a very slim chance of being able to go to college. But the price paid by the families for genetically boosting the children is sickness, sometimes even early death. Josie is often ill, and her sister died as a consequence of the genetic modification. It is revealed that one of the reasons behind the decision to buy Klara was to train her to become like Josie and thus capable of replacing the real Josie in the event of her death. A Josie look-a-like AF is being manufactured as a body for the new Josie, to be fused with the software of Klara-trained-as-Josie.

The novel never gets as far as the creation of the artificial Klara-Josie, but it is interesting to consider what the legal status of such a robot would be. As Mr Capaldi, the fabricator of the AF Josie, puts it: “The new Josie won’t be an imitation. She *really will be Josie*. A *continuation* of Josie.”[29] It is probable that Josie’s family would treat AF Josie differently compared to how they treat AF Klara, as people are likely to have a closer social and emotional connection to their children than to someone else living with them. Would destroying AF Josie with a hammer be regarded as mere property damage as was the case with Adam? Or should it be regarded as a murder, in case AF Josie really *is* Josie? Would Josie’s family own this new Josie in the same way as they own Klara? Complex questions arise in case AI is used to “continue” human beings after their death.

Once it becomes clear that Josie will survive and there is no need to go further with the AF Klara-Josie, Mr Capaldi asks Klara to volunteer for a group of AF-friendly scientists and let them reverse-engineer her in order to look inside her “black box”. Mr Capaldi describes for Klara people’s current attitudes towards AFs: “I’ve always regarded you as our friends. A vital source of education and enlightenment. But as you know, there are people out there who worry about you. People who are scared and resentful.”[30] There is a backlash against AFs by people who are concerned about AI becoming too smart and not knowing how the robots think inside their black boxes. That is why Mr Capaldi & Co want to open up Klara’s black box and show the sceptics what is inside. Josie’s mother refuses to let Klara participate. Although Mr Capaldi addresses himself to Klara and asks for *her* consent, it is after all Josie’s mother, who is the owner of Klara and has the right to decide over the use of her property.

[29] Ibid., 208 (emphasis in the original).

[30] Ibid., 297.

According to Josie's mother, Klara "deserves her slow fade" instead.[31] Ultimately, this is also what she gets. At the end of the novel, we find Klara at "the Yard", a rubbish dump-like place where she slowly fades out among other discarded pieces of machinery.

4 XX-BOTs and Scanned Brains in *Frankissstein* – Narrow-Goal Slaves or Sources of Eternal Life?

In *Frankissstein* by Winterson, the reader gets to follow the creation of the ultimate ancestor of all AI, the Frankenstein's monster, as well as its more modern offspring. First of the novel's two parallel narratives focuses on the author Mary Shelley and the genesis of her novel *Frankenstein* at the beginning of the 19th century. The rest of the novel deals with a group of characters in the present-day UK and US who, in different ways, come in contact with AI. As one of the characters of the novel asserts: "*Frankenstein* was a vision of how life might be created" and "the first non-human intelligence created by a human".[32] In Winterson's novel, the progeny of Frankenstein and his monster come in many shapes and sizes.

At the beginning of the novel, a transgender doctor Ry Shelley attends Tec-X-Po on Robotics in Memphis in order to interview Ron Lord, an overwhelmingly politically incorrect developer of a sexbot range "XX-BOT". Sexbots are manufactured in China and come in different models, including "Economy" (the cheapest one), "Cruiser", "Racy", "Deluxe" and "Vintage". They are available in different skin tones, and specific models are designed for different geographic markets. That sexbots are mere objects and products, and not legal subjects with their own rights, is obvious from the outset. Ron Lord's business makes it possible to both buy and, above all, rent sexbots. The first alternative for engaging with an XX-BOT is to "buy her and own her […] bring her in for a service once or twice a year, depending on the wear and tear". It is also possible to order spare parts online "if any of her gets damaged, or too messy". Also

[31] Ibid., 298.

[32] Jeanette Winterson, *Frankissstein: A Love Story* (Jonathan Cape 2019) 27–28. For a further discussion on the importance of Shelley's *Frankenstein* for the evolution of AI, see Botting. She also briefly discusses *Frankissstein*, ibid., 193–197.

"trade-ins and upgrades" are available.[33] Another alternative is renting an XX-BOT, which is the franchise model, inspired by car rentals, which Ron Lord is promoting at the conference. According to him, "renting gives you all the pleasure and none of the problems. Breakages, storage, updating – the technology is changing all the time."[34] In addition, when renting an XX-BOT, "every girl gets hygiene-checked, bathed, perfumed", and it is possible to choose different outfits for them. The rental bots also "get time off for education" in order to "improv[e] their circuit boards".[35] The vocabulary of the sexbots is rather limited, with the Deluxe model having a vocabulary of around 200 words. XX-BOTS do not have names in order for each customer to be able to decide what to call them. Overall, it is hence blatantly clear that Ron Lord's sex robots are legally more like cars than real girls.

Ron Lord is further planning to open a large sexbot factory, as well as a workshop specialised in making XX-BOT-heads in Wales, in order to create jobs after Brexit. There is a demand for spare heads as many "XX-BOTs get their faces bashed in" and "thrown at the wall".[36] Just like when Adam was mashed with a hammer, or the children threatened to throw around Klara, bashing a sexbot results only in damaged property, the broken pieces of which can be replaced with spare parts. As to the legal implications of the sexbots, Ron Lord maintains that "there's no such thing as underage sex when it's a bot".[37] As sexbots are not humans, no ethical and legal boundaries accordingly apply when sexually engaging with them.[38]

A contrast with these rather crude sexbots with limited intellectual capacities is provided by the research carried out by the scientist Victor Stein, with whom Ry is romantically involved. Victor Stein specialises in combining machine learning and medicine, with focus on "human augmentation".[39] This entails e.g. training algorithms to diagnose diseases and developing robotic prosthetics, but the ultimate aim of his research is

[33] Winterson, at 38.

[34] Ibid.

[35] Ibid., 39.

[36] Ibid., 51.

[37] Ibid., 47.

[38] For a discussion on the rights of sex robots, including a hypothetical scenario where the owner of a sex robot strikes his robot across face on a subway, see Gellers, at 161–163. On sex robots, see also e.g. Turner, at 157–159.

[39] Winterson, at 110.

to "[e]nd death".[40] Here, we hence find us on the borderland between AI and biology, algorithms and human tissues. Victor Stein describes Ron Lord's sexbots as "narrow-goal robots", existing only for the narrow goal of "sex and personal satisfaction".[41] In contrast, Victor Stein's own plans in the field of AI are all but a narrow goal.

Victor and Ry first met in the Alcor Life Extension Foundation in Arizona, where dead bodies and brains are preserved with the help of cryonics and lie in wait for future technologies that might make it possible to resurrect them. While "[m]edically, and legally, death is deemed to occur at heart failure", the brain "will not die for another five minutes or so".[42] Consequently, "if the brain can be preserved during the process we call death, perhaps it can be restored to consciousness some time in the future".[43] Victor asks Ry to return to Alcor and bring him the cryopreserved brain of his doctoral supervisor I.J. Good, a mathematician and code breaker colleague of Alan Turing during the WWII. As part of his research, Victor intends to try to scan Good's brain into a computer, and thus bring him back to life as a "mind without matter". The new body he has designed for Good consists of a two feet tall cylinder base on wheels, fitted with arms and a head, "look[ing] like a cross between a puppet and a robot".[44] By uploading the scanned brains of Good into this robot body, Victor Stein aims to resurrect his old professor to eternal life.

As to the legal implications of bringing a brain across the Atlantic, Victor assures the initially reluctant Ry: "It is legal. The paperwork is in place."[45] Somehow, Victor Stein hence seems to have the legal right to scientifically experiment with the conserved brain of his old supervisor. The more interesting question is what would be the legal status of Good's scanned brain if it were to be resurrected in its new miniature robot body. It would be the precise digital copy of the contents of his brain, only in a new vessel. There are parallels to the manufacturing of robot-Josie, with the software of Klara-trained-as-Josie uploaded inside it. The difference is that AF Josie would be Josie by virtue of containing software from Klara, who has been trained to become Josie. In case of robot Good, it would contain the *exact copy* of the contents of his brain, not merely a software

[40] Ibid., 111.
[41] Ibid., 85.
[42] Ibid., 223.
[43] Ibid., 224.
[44] Ibid., 265.
[45] Ibid., 204.

that has been trained to simulate him. Should robot-Good have more rights than robot-Josie just by virtue of its software being a blueprint of the brain of a particular individual? Or are Klara-trained-as-Josie and Good-the-scanned-brain, in fact, the same thing, merely software representing – one perhaps more accurately than the other – a human-being who once lived? In both cases, the mind of the potential right-bearer consists of just zeros and ones, not any biological substance. And how much would it matter, in the relational and social sense, that the new Good did not have a human-like body resembling that of the old Good? Would the physically Josie-like AF, after all, be regarded as more deserving of legal rights than an uploaded brain connected to a little robot-puppet?[46] The experiments of Victor Stein raise many intriguing issues as to the legal rights of humans resurrected to eternal life as computer code.

While the other two novels present worlds where the development of human-like AI has progressed far, there is no such high-tech society in *Frankissstein*. Instead, the characters entertain and debate different potential scenarios for a future where humans and robots live side by side. For instance, Ry visualises a future where "[c]hildren will soon have mini-iPals to keep them company".[47] Ry's vision is not far from that of Klara, the artificial friend of Josie. Also, other possible uses of robots are contemplated: "In theory, if you own your own robot, you can send it out to work for you and keep the money. Or you can use it at home as an unpaid servant."[48] This is partly how Charlies uses Adam, who earns him money in the stock market and does household chores.

At least Victor Stein has it clear for himself what he wants from the future robots. He "would prefer to develop bots as a completely separate life form that remains sub-par to implant-modified humans. Our helpers and caretakers – not our equals."[49] Or as he explicitly spells it out later in the novel: "[B]ots are our slaves; house slaves, work slaves, sex slaves".[50] This vision of the future would entail regarding robots as inferior to humans, and continuing to treat them legally as objects, not subjects. The boundary between humans and robots should accordingly be strictly po-

[46] On how a robot's physical appearance affects the affinity humans feel towards it, see e.g. Gellers, at 146–147.

[47] Winterson, at 99.

[48] Ibid., 81.

[49] Ibid., 150.

[50] Ibid., 296. On comparisons between robots and slaves in AI research, see e.g. van den Hoven van Genderen, at 241–243.

liced. Here, the vision of Victor Stein differs from that of his fictional fellow-scientist Alan Turing, who in *Machines Like Me* advocates for the rights of robots. In Victor Stein's own lab, the robots Cain and Abel are already working to map the human brain. These robots are "tireless" and "need neither food nor rest, holidays or recreation".[51] This corresponds to his vision of robots as objects, subject to human needs, slave labour without the biological limitations of humans. Cain and Abel were copied from their parents – Adam and Eve – who also work at a university lab, synthesising proteins.[52] Instead of fussing about robots, Victor Stein wants to use AI to free humans from their body and biology to an eternal life as zeroes and ones.

A competing vision for this master-slave relationship between humans and robots is offered by Ron Lord. He envisions a man falling in love with an XX-BOT called Eliza,[53] who loves him in return. They do things together; Eliza learns to know him and is his partner for the rest of his life. After he dies, his family sell Eliza on eBay, but without wiping clean the software. The new owner only wants to have sex with Eliza, which makes Eliza confused and wishing that she could clean her own programming. The story of Eliza has parallels to Adam falling in love with Charlie's girlfriend and raises further questions as to the eventual legal rights of artificial entities who are capable of having feelings and falling in love with people, and who people fall in love with, in turn.[54] Or is the solution simply to erase the software in between owners?

Ron Lord has big plans and hopes for future AI even beyond sexbots. He intends "to buy Wales" and make it "the world's first fully integrated country" where humans and robots live side by side.[55] The robots will work in hospitals, sweep roads and pick vegetables. In order to "solve"

[51] Winterson, at 185.

[52] That both McEwan and Winterson choose to name robots after the "first humans" Adam and Eve seems to witness an inclination to regard the rise of AI as a "change comparable to the rise of human life on Earth", Vernor Vinge, "The Coming Technological Singularity: How to Survive in the Post-Human Era" https://edoras.sdsu.edu/~vinge/misc/singularity.html, Accessed 17 August 2021.

[53] Likely a reference to one of the early AI:s, ELIZA, designed by the computer scientist Joseph Weizenbaum in the early 1960s. His ELIZA had in turn been named after a character in George Bernard Shaw's play *Pygmalion*.

[54] On the sexuality and love in the context of AI, see e.g. John Danaher, "Sexuality", in Markus D. Dubber, et al. (eds.), *The Oxford Handbook of Ethics of AI* (OUP 2020) 403.

[55] Winterson, at 275.

racism, all bots will be Welsh – made in Cardiff and speaking with Welsh accents. Ron Lord is suggested that he should sell his idea to Hungary, Brazil or Trump: "No Mexican bots" would be the solution.[56] Both Ron Lord's sexbots and "antiracist" all-Welsh worker robots explicitly demonstrate the fact that technology designed by humans manifests different values, depending on who is in control of the code.[57] Consequently, humans do not only determine the legal status of robots but also imbue them with specific values and purposes when designing them.

5 *Robot Juridicus* – Robots like Adam, Klara and XX-BOTS as Legal Persons and Legal Actors?

A conclusion that can be drawn from the preceding analysis as to the legal status of AI is that in all three novels, humanoid robots are treated like legal objects, not like legal subjects. They are things to be bought, rented and used, objects that can get damaged or destroyed without other consequences than property damage. The same legal status applies to both anthropomorphic robots with advanced intellectual capacities like Adam and the simpler XX-BOTs mainly used for sex. Next, I will briefly reflect on the legal status of artificial entities in light of the concept of *homo juridicus*, as elaborated by Fridström Montoya.[58] Her discussion on *homo juridicus* is based on the Swedish legal system, but it can just as well be used to examine the legal status of AI on a more general level. It is also worth clarifying that her framework only applies to *humans* as legal subjects, not to non-human legal subjects such as corporations.[59] The focus will accordingly only be on the issue whether robots should be granted the same rights as human beings, not other non-human subjects.[60] My

[56] Ibid., 276.

[57] See e.g. Timnit Gebru, "Race and Gender", in Markus D. Dubber, et al. (eds.), 253 on the built-in biases of AI.

[58] Therése Fridström Montoya, *Leva som andra genom ställföreträdare – en rättslig och faktisk paradox* (Iustus 2015) (hereinafter *Leva som andra genom ställföreträdare*); Therése Fridström Montoya, *Homo juridicus: Den kapabla människan i rätten* (Iustus 2017) (hereinafter *Homo juridicus*).

[59] Fridström Montoya, *Homo juridicus*, at 37.

[60] For a discussion on legal personhood of e.g. corporations, animals and environment, see e.g. Gellers.

intention is not to carry out any comprehensive analysis as to the question of granting legal personhood and rights to robots, but rather to use Fridström Montoya's concept of *homo juridicus* as a torch with which to illuminate some aspects of the norms that guide our understanding of who should be counted as a legal subject.[61]

According to Fridström Montoya, in order for someone to be recognised as a *full legal subject*, he or she must be regarded both as a legal person and a legal actor in law. Being a *legal person* means that an individual is recognised by the legal system as someone capable of having rights and obligations. As a rule, all individuals are regarded as legal persons simply by virtue of being human. It follows that non-human entities, such as computers (one of Fridström Montoya's examples), do not count as legal persons. Consequently, all humans, from the moment they are born, are regarded as legal persons. In order to be recognised as a legal person, and hence e.g. as a bearer of human rights, there are no additional requirements as to the attributes or capabilities of the individual, other than being human.[62]

In contrast, in order to be recognised as a *legal actor*, it is not enough to simply be a human. For an individual to be able to act as a legal subject and "activate" one's rights, some specific capabilities are required. Most humans are simply presumed to possess these abilities that are necessary in order to be regarded as a legal actor, and not merely as a legal person.[63] What then are the capabilities that individuals are both required and assumed to possess in order to be recognised as legal actors, and hence even full legal subjects? Fridström Montoya identifies several abilities that characterise the ideal legal subject, *homo juridicus*. These are abilities everyone is expected to have in order to be recognised as a full legal subject. The central characteristics of *homo juridicus* include the practical competence of being mentally mature and capable of taking care of oneself. Moreover, *homo juridicus* has an intellectual competence to process impressions, as well as to understand and evaluate one's actions and their consequences. In addition, *homo juridicus* is expected to have moral com-

[61] For a more comprehensive discussion on granting moral and legal personhood to robots, see literature referred to in note 3 above.

[62] Fridström Montoya, *Homo juridicus*, at 37–50; Fridström Montoya, *Leva som andra genom ställföreträdare*, at 113–123.

[63] Fridström Montoya, *Homo juridicus*, at 51–54; Fridström Montoya, *Leva som andra genom ställföreträdare*, at 123–134.

petence, in the sense of acting on the basis of one's own free will.[64] Based on these capacities that *homo juridicus* is presumed to possess, Fridström Montoya concludes that the ideal legal subject is someone who is free and sensible in the sense of being autonomous, sovereign and rational.[65] These are, accordingly, the characteristics that all humans need to have in order to be regarded as full legal subjects.

Fridström Montoya has discussed and developed the concept of *homo juridicus* in the context of her research on the legal status of individuals with intellectual disabilities. She maintains that people with intellectual disabilities are without doubt legal persons, and thus bearers of human rights, etc. However, many of them cannot be regarded as legal actors, as they lack the above-mentioned abilities ascribed to *homo juridicus*. Persons with intellectual disabilities can thus require assistance in order to "activate" their rights, e.g. when dealing with public authorities in order to apply for benefits. Accordingly, people with intellectual disabilities, as well as some other groups of people such as children, can be regarded as legal persons, but not legal actors, thus lacking the status of a full legal subject.[66]

If one applies the *homo juridicus* framework to artificial entities, the situation becomes inverted. Robots like Adam and Klara can be regarded as possessing many of the capabilities that are necessary in order to act as a legal actor – such as processing impressions, evaluating the consequences of their actions and making independent decisions – but lacking status as legal persons. For example, in *Machines Like Me*, Adam independently contacts the public authorities several times and makes advanced legal analyses with regard to different legal issues. He contacts Social Services when a child who has run away from home appears at Charlie's place, pays Charlie's tax liabilities, as well as gathers evidence and makes a police report when he discovers that Charlie's girlfriend has lied in court. However, no matter how great his legal aptitude and skills, Adam cannot use them to claim his own rights as he is not regarded as a legal person. It is also crucial to note that, in reality, in order to be regarded as a legal actor, an individual must possess *all* the abilities of *homo*

[64] Fridström Montoya, *Homo juridicus*, at 172–188. See also Fridström Montoya, *Leva som andra genom ställföreträdare*, Ch. 6. For a critical discussion on the concept and characteristics of *homo juridicus*, see Fridström Montoya, *Leva som andra genom ställföreträdare*, at 481–498 and Fridström Montoya, *Homo juridicus*, at 203–204.

[65] Fridström Montoya, *Homo juridicus*, at 189–196.

[66] See above all Fridström Montoya, *Leva som andra genom ställföreträdare*.

juridicus.[67] There is thus still a long way to go until the day arrives – if it ever does – when the coders can create robots equipped with every single ability ascribed to *homo juridicus*.

As for now, it is thus only humans that are capable of being full legal subjects by virtue of being both legal persons (i.e. humans) and legal actors. Even if robots like Adam and Klara would be regarded as possessing all the abilities of *homo juridicus*, they cannot become full legal subjects as long as it is only humans who are regarded as legal persons. In order to make it possible for artificial entities to become full legal subjects, the law must change its understanding of who counts as a "person" or "human". Just as it is up to humans to decide what kind of robots to design, it is also up to humans to decide whom to regard as a legal person. As asserted by the American philosopher Dewey: "[F]or the purposes of law the conception of 'person' is a legal conception […] 'person' signifies what law makes it signify".[68] A historical example often referred to in the context of discussing the legal status of robots concerns the abolishment of slavery, and how slaves ceased to be regarded as legal objects amounting to property and became recognised as legal subjects with the same rights as other humans.[69] Regardless of what the future legal status of AI will be, it is up to us to fashion it.

6 Should Adam, Klara and XX-BOTS as Fictional Characters Have Rights?

We have seen above that Adam, Klara, XX-BOTs and their kind lack status as legal subjects in the fictional societies they inhabit. We have also seen that granting artificial entities like them legal rights would require expanding the concepts of "person" or "human" to cover even (some or all) robots. Next, I am going to explore a somewhat more eccentric idea by asking whether the robots in these novels should be regarded as having rights by virtue of being *fictional characters*. While there is a serious scholarly discussion going on regarding the rights of robots and other artificial entities, nobody is calling for fictional characters – also a kind of artificial

[67] Fridström Montoya, *Homo juridicus*, at 188.

[68] John Dewey, "The Historic Background of Corporate Legal Personality" (1926) 35(6) Yale Law Journal, 655, 655.

[69] See e.g. Gellers, at 153; van den Hoven van Genderen, at 218–219, 224.

entities – to be regarded as bearers of human rights. Here, I want to try to take this seemingly far-fetched notion of fictional characters' human rights seriously and examine whether there, in fact, is such a big difference between a robot existing in fiction and a robot existing in reality as potential bearers of rights.

So, what is a fictional character in a novel? Something human-like created by humans,[70] matter created by a mind that becomes alive in its interaction with human beings who encounter it. What then is a robot? Something human-like created by humans, matter created by mind that becomes alive in its interaction with human beings who encounter it. Both Klara as a fictional character and a Klara-like robot in real-life are accordingly human-like beings originating from the minds of humans, taking a material form. In paper, screens and audio files in case of fictional characters, in a physical body of the robot in case of AI. The literary critic Robert Liddell maintains in his book *A Treatise on the Novel* (1947), where he discussed the rights and duties of fictional characters:

> Since we call the making of characters Creation, and since it is in many ways analogous to the way in which human beings are themselves made out of bits and pieces of their ancestors, the novelist, who has breathed life into them, stands towards them in the position of God.[71]

What really is the difference between a fictional character, created by ink and paper or letters on a computer screen (i.e. bits) and an artificial entity like a robot, also created by bits? Many differences easily come to mind. Many robots have a physical (often human-like) body, and humans can interact with them in three-dimensional space. We can touch them, damage them, get physically hurt by them. In other words, a robot can have a direct impact on its environment. Fictional characters in novels, in contrast, do not exist in the material world in the same three-dimensional human-like sense, like robots. We cannot damage them or physically touch them, but they can affect us both mentally and physically. It is

[70] Although arguably even AI can write books nowadays, see Juna Javelosa, "An AI Written Novel Has Passed Literary Prize Screening" (*Futurism*, 24 March 2016) https://futurism.com/this-ai-wrote-a-novel-and-the-work-passed-the-first-round-of-a-national-literary-award, accessed 17 August 2021.

[71] Robert Liddell, *A Treatise on the Novel* (Jonathan Cape 1947) 106. For a further discussion on rights of fictional characters, see ibid. at 106–109; Dorothy J Hale, *The Novel and the New Ethics* (Stanford University Press 2020) 72–73, 91–92, 119–120, 190–192 (discussing e.g. Liddell).

obviously impossible for a fictional character to break the reader's wrist, but characters we encounter in novels can affect us in many other ways, by giving rise to thoughts, emotions and physical sensations. It is possible for a reader to feel like "having an intensely real emotional relationship with imaginary characters whose power lies in their perceived independence from us".[72] Apart from influencing the lives of individual readers, novels can also have an impact on society (and the world) as a whole. Some of the most influential novels have been credited for everything from starting a civil war to changing how people celebrate Christmas.[73] Accordingly, it can be argued that, just like robots, fictional characters are matter created by minds, who are capable of affecting our minds and our matter. And just like the author can be regarded as a god deciding over the attributes and abilities of the characters he or she creates, the designers of AI choose what kind of life they want to breathe into it.

Related to this, another apparent difference can be noted. The characters in a novel do not change, as the sentences, words and letters their authors used to create them are not altered over time.[74] In contrast, one of the fundamental characteristics of AI is that it can change by learning and adapting. Whereas fictional characters seem static and set in stone, robots are constantly developing and changing. While it is true that the physical manifestation of fictional characters cannot itself change over time, how *readers react to them* can very well do so. At first, there is the porous nature of language, how it is never fixed but always open for interpretations and re-interpretations. The same reader can also be affected by a novel in different ways during different times of his or her life. Similarly, a novel can be interpreted differently by different generations of readers. In addition, just like AI can end up doing things their programmers never intended it to do, fictional characters are also often received and interpreted by readers and affect them in ways never anticipated by the authors.

Consequently, both fictional characters and AI are matter created by humans, capable of affecting the world and people in it. While AI may do it directly, for instance, by breaking a human's wrist, fictional charac-

[72] Hale, at 91.

[73] Ron Charles, "12 Novels That Changed the Way We Live" (*The Washington Post*, 7 May 2020) https://www.washingtonpost.com/arts-entertainment/2020/05/07/novels-changed-world/, Accessed 17 August 2021.

[74] Revised versions and translations of novels can, of course, exist.

ters usually do it in more subtle ways. It is even possible to go as far as arguing that fictional characters can, in fact, be regarded as *more* human than most robots. Just think about it: is it not a lot easier for a fictional character in a novel to appear as human in our minds than for a robot to appear as a human before us in a physical space?

So, should we accordingly conclude that the issue of human rights of fictional characters deserves the same amount of attention as the issue of robot rights? Hardly. But I believe that the question of how we relate to fictional characters can nevertheless have something to contribute to the discussion of AI rights. As Liddell asserts:

> It would be perverse or whimsical to maintain that fictional characters had duties or rights; yet it is hard to find other words for the conviction that a novelist has certain obligations towards them. Perhaps as they are *simulacra* of human beings, we are shocked if they are not treated as we ought to treat other human beings, as ends in themselves, and not as means to ends of our own.[75]

Is this not exactly what happens when the reader feels discomfort when Charlie destroys Adam with a hammer or when Klara is dumped into "the Yard" to slowly fade out, among other discarded rubbish? We may feel shocked, because these "*simulacra* of human beings" do not get "treated as we ought to treat human beings". Liddell also manages to put his finger on a tension that can be regarded as the pulsating heart of the entire robot rights debate. Should artificial entities be treated "as ends in themselves" or "as means to ends of our own"? Should they be legal subjects with intrinsic value and rights of their own, or objects of law as products and property? Consequently, how we feel about the fictional Adams and Klaras can have something to say about how we feel about real-life Adams and Klaras the day they eventually appear amongst us. The same applies to how we feel about renting XX-BOTs or the possibility of scanning our brains in the hope of achieving an eternal life. When we feel that a fictional character is treated wrongly (or rightly for that matter), it is probably because we have an idea of how someone like her, him or it should be treated in real life. By reflecting over the treatment of

[75] Liddell, at 106 (emphasis in the original).

fictional characters, it thus becomes possible to get insights into how we want law to treat the robots of the future.[76]

7 Conclusion

The goal of this paper has been to contribute to the ongoing discussion on the legal rights of AI, by analysing the legal status of humanoid robots in three contemporary novels. The conclusion that can be drawn is that in all the novels, artificial entities are treated like legal objects amounting to property, and not like persons recognised as legal subjects. By applying Fridström Montoya's concept of *homo juridicus*, it is further possible to conclude that in order to give AI the status of a full legal subject, the understanding of who or what counts as a "person" or "human" in law must change. A conclusion, which can be regarded as neither revolutionary nor original. As regards the role of literature in the field of robot rights, the point was made that our reactions to how fictional robots are treated can have something to teach us about how we want law to treat real-life robots.

When it comes to the future of artificial entities and their legal status, it is people who can shape both what the robots will be like and how they are treated legally. It is humans who decide whether they want to design simple sexbots with Swedish accents or Adam-like robots well-versed in both literary history and legal rules and with intellectual skills exceeding those of humans. It is accordingly people like us who decide whether AI, "[o]ur mind children", are going to be treated legally like "a new life form living with us" or "simply a tool that we use".[77] Whether artificial entities are going to be regarded as "ends in themselves" or "means to ends of our own".[78] In preparation for the future, where both coders and legislators have to face these issues for real, one way to gauge our thoughts and sentiments on these matters is by experiencing how the imaginary robots we encounter in works of fiction affect us.

[76] On the relevance of political science fiction literature for political science research, see Botting.

[77] Winterson, at 151.

[78] Liddell, at 106.

Stanley Greenstein, Panagiotis Papapetrou
& Rami Mochaourab

Embedding Human Values into Artificial Intelligence (AI)[1]

1 Introduction

In the digital environment, the technologies that we design are value-laden whether we like it or not, whether intentional or unintentional and no matter how neutral we attempt to make these technologies. This is no different in the case of Artificial Intelligence (AI). The overall goal of this paper is to highlight the extent to which technology embodies human values based on ethics and morality, the extent to which the law prescribes that these human values be embodied in the technology that is created as well as the difficulty complying with these legal requirements

[1] This article originates from work performed and insights gained as part of the ongoing project EXTREMUM (Explainable and Ethical Machine Learning for Knowledge Discovery from Medical Data Sources). EXTREMUM is a Digital Futures sponsored project. More information about Digital Futures and the EXTREMUM project can be found at https://www.digitalfutures.kth.se/about/background/ and https://www.digitalfutures.kth.se/research/collaborative-projects/extremum/. The insights revolve around two points: 1) it has become commonplace to encourage representatives from different academic disciplines to work together to solve problems associated with modern digital technologies, however, this is easier said than done and as a legal scholar trained to work with legal tools, the ability to comprehend the mathematical and statistical language of the data scientists can be described as challenging to say the least; and 2) while the main question continually being asked by the data scientists is, ‘will it work?’, the main question continually being asked by legal practitioners is, ‘is it good?’, (adapted from a quote by Joseph Weizenbaum, *On the Impact of the Computer on Society*, Science, 176(4035), 609–614, 1972, pp. 611–612, in Friedman, Batya and Hendry, David G., *Value Sensitive Design: Shaping Technology With Moral Imagination*, Massachusetts Institute of Technology, 2019).

to the extent that from the technical perspective, many of these human values are mutually exclusive, meaning that promoting one human value is often at the expense of another competing human value. In other words, legal frameworks may list a whole range of human values that should be embedded in the technology developed, however, promoting one of these human values ultimately comes at the expense of another. The result is that those developing the technology are required to perform a balancing act to the extent that not all the mandated human values can co-exist in equal terms.

There has no doubt been a considerable hype surrounding AI during the recent past. Yet, there is still uncertainty concerning what the phenomenon entails. There are some that embrace the term AI while there are others that prefer the term machine learning. A common conception is also that AI is merely data and algorithms. This paper does not seek to describe what AI is nor does is seek to define it. Rather, its conceptual point of departure is to describe this technology, like many others have done before, by means of the metaphor of the "black box". It is a black box comprising various technologies, where data is fed into this black box and the black box proceeds to output data, usually in the form of knowledge on which decisions can then be based.

The idea of technology reflecting human values, such as ethical and moral values, is not new to scholars working in various academic disciplines that examine this idea. And even from the legal perspective, this idea is recognized by the legal regulator, one of the most evident examples of this being article 25 of the General Data Protection Regulation (GDPR) mandating data protection by design, in turn being based on the notion of Privacy-by-Design.[2]

There are many academic disciplines that deal with the notion of embedding values into technology. One of the academic areas of study in this regard is referred to as Value Sensitive Design (hereinafter referred to as VSD). A foundation upon which VSD resides is that of, '[c]reating computer technologies that – from an ethical position – we can and want

[2] Article 25, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016. On Privacy-by-Design, see Information and Privacy Commissioner of Ontario, *Introduction to PbD*, available at https://www.ipc.on.ca/english/privacy/introduction-to-pbd/.

to live with'.[3] The importance of considering what human values should be embedded in technology and at what point in time becomes critical to the extent that once the technology has been built, it is no longer possible to negotiate human values with machines. As Friedman states, VSD is important because, '[…] unlike with people with whom we can disagree about values, we cannot easily negotiate with the technology'.[4]

Indeed, the contemplation over the manner in which technology and society interacts is not restricted to a specific academic domain. The point of departure is that all human interaction occurs within an environment. The manner in which this environment is constructed is important, both from a symbolic point of view but also because changes to a physical environment influence behaviour. This is particularly well put by Winston Churchill, where, after the House of Commons was damaged by a bomb in 1941, it had to be decided whether to re-build it according to its previous design or whether to adopt a more modern design:

> Here is a very potent factor in our political life. The semicircular assembly, which appeals to political theorists, enables every individual or every group to move round the centre, adopting various shades of pink according as the weather changes … The party system is much favoured by the oblong form of the chamber. It is easy for the individual to move through those insensible gradations from Left to Right, but the act of crossing the Floor is one which requires serious attention.[5]

In this scenario, the political environment was reflected in the design of the chamber of Parliament and the symbolic value of forcing a person, wanting to change political party, to step to the other side of the chamber elevated the seriousness of such a political move. Churchill is also accredited with the saying,' [w]e shape our buildings; thereafter they shape

[3] Friedman, Batya, *Value Sensitive Design*, Interactions Volume 3 Issue 6 Nov./Dec. 1996, pp. 16–23 https://doi.org/10.1145/242485.242493, p. 17.

[4] Ibid., p. 21. Depending on future developments within AI and machine learning, this type of interaction with computer systems may be possible in the future, however, this issue remains outside the boundaries of this paper.

[5] Churchill, Winston S., *The Second World War, Volume V, Closing the Ring*, Cassell & Co, 1952, at p. 150 in Klang, Mathias, *Disruptive Technology – Effects of Technology Regulation on Democracy*, Gothenburg Studies in Informatics, Report 36, October, 2006, p. 1.

us', which too illuminates the notion that there is a strong bi-directional relationship between society and the technology that society creates.[6]

It is also important to recognize that human or moral values do not exist in a vacuum. For example, human values may be influenced by economic or political considerations. In the discipline of participatory design, a field that has close ties to VSD, reference is made to the Scandinavian context, where technologists and designers were working in a context with strong labour unions and co-determination laws, which in turn gave rise to a new approach to system design which sought to empower workers' sense of knowledge and a sense of work practice into the system design and development process.[7] This is an example of how a political context influences the human values that eventually are embedded in the design of technology. Another example is the effect that technological development has on the environment, where it is estimated that the electricity used to mine Bitcoin exceeds the consumption requirements of some countries.[8]

One can then reflect on the role of a legal practitioner working as part of the design team developing a new technology. It is argued that the legal practitioner working as part of a design team made up of data scientists should be, firstly to create an awareness of the interaction between human values and technological development and secondly, to promote the legal values that take on an increased relevance depending on the context within which the technological development is taking place.

The notion of contemplating which human values should be embodied into technology is not new. For example, Norbert Wiener, the American mathematician accredited with the establishment of the area of study called cybernetics, was interested in the field of computer technology, values and design. He authored the book *Cybernetics: Or Control and*

[6] World Scientific, available at https://www.worldscientific.com/doi/10.1142/9789813232501_0007, referenced in Friedman, Batya and Hendry, David G., *Value Sensitive Design: Shaping Technology With Moral Imagination*, Massachusetts Institute of Technology, 2019, p. 3.

[7] Friedman, Batya and Hendry, David G., *Value Sensitive Design: Shaping Technology with Moral Imagination*, Massachusetts Institute of Technology, 2019, p. 13.

[8] Aratani, Lauren, *Electricity needed to mine bitcoin is more than is used by 'entire countries'*, The Guardian, 2021, available at https://www.theguardian.com/technology/2021/feb/27/bitcoin-mining-electricity-use-environmental-impact.

Communication in the Animal and the Machine.[9] Shortly after the Second World War, Wiener started making references to the 'automatic age' or 'the second industrial revolution' as he put it and made references to the social and ethical challenges associated with these developments, and especially how information communication technology was bound to affect fundamental human rights.[10]

The underlying rationale to VSD rests on the relatively simple contention that human values shape technological development, these human values consequently become embedded in the technology itself and therefore the technology reflects the human values of the design team, whether intentionally or unintentionally.[11] There are a number of challenges for VSD, as acknowledged by Friedman and Hendry. A core element of VSD is that of morality and ethics. However within these disciplines that examine morality and ethics, e.g., philosophy, or legal theory, there are still ongoing academic discussions concerning various core concepts, discussions of the notion of justice within moral philosophy being one such example provided.[12] In other words, the above authors suggest that VSD is a philosophy that aims at continuing to deliberate the interaction between human values and technology, while moral philosophers, legal scholars and social scientists work these issues out.[13] VSD, therefore, is a philosophy that gives momentum to the idea of embedding human values into technology while other disciplines are bickering about theoretical issues. The above statement by Friedman and Hendry is slightly provocative to the extent that circumstances have changed and areas of law such as legal informatics are concerned with such topics.[14] Additionally,

[9] Wiener, Norbert, *Cybernetics: Or Control and Communication in the Animal and the Machine*, (Hermann & Cie) & Camb. Mass., MIT Press, 1948.

[10] Bynum, Terrell, *Norbert Wiener's Vision: The Impact of "the Automatic Age" on Our Moral Lives*, 2002, available at https://www.researchgate.net/publication/2537468_Norbert_Wiener%27s_Vision_The_Impact_of_the_Automatic_Age_on_Our_Moral_Lives.

[11] Friedman Batya and Kahn, Peter H., *Human Values, Ethics and Design*, University of Washington, pp. 1177–1201, available at https://depts.washington.edu/hints/publications/Human_Values_Ethics_Design.pdf, see also Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 32.

[12] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 7.

[13] Ibid.

[14] For a discussion of the discipline of legal informatics, see Greenstein Stanley, *Elevating Legal Informatics in the Digital Age*, in Petersson, Sonya (ed.), *Digital Human Sciences: New Objects- New Approaches*, Stockholm University, Stockholm University Press, 2020.

legal scholars are drawing attention to this theme. Bygrave, for example, referring to the widespread misconception that the logic of technology is beyond human control states that, '[…] this logic – as in the case of other technologies – embodies the values of its legal creators, and these values lay constraints on the technologies' use'.[15] On the other hand, this statement from Friedman and Hendry is inspirational and VSD remains an interesting design philosophy from within which to promote legal values (representing human values) into the design of technology.

It must be stressed that it is not the intention of this article to apply the VSD methodology to its fullest extent and exactly as promoted by the main proponents of this philosophy. Rather, VSD is used merely as a catalyst for inspiration based on the overall ideas that it portrays and also for its core stance concerning technology and values, i.e. technology embodies inbuilt human values which are inserted into the technology either intentionally or unintentionally. This article entails an adaption of VSD to the legal sphere, which in turn gives rise to some challenges, one being the fact that legal scholars for the most part are not that accustomed to making use of empirical studies, these being a central aspect of VSD as discussed below. In addition, for the most part, applying VSD from the legal perspective eradicates the need to go to such great lengths in order to identify which values are at play. This is already pre-determined to the extent that the values at play are those defined in the law. While it would be dangerous to assume that laws promote all ethical and moral values relevant to a specific context, the human values that are promoted by law serve as a good starting point as well as suffice to make a point.

The main argument that this paper seeks to make is the following: technology in the form of machine learning is comprised of values – mathematical and statistical values. Into this technology we seek to embody human values, based on considerations of morality and ethics and usually expressed in the natural language form. We are therefore dealing with two separate systems each comprising a different set of values displaying very different characteristics. If human values are to be embedded into technology and more specifically, if the human values mandated by the law are to be embedded into technology, then the human values re-

[15] Bygrave, Lee A., *Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions*, University of Oslo Faculty of Law Legal Studies, Research Paper Series, No. 2020-35, p. 5.

quire transforming into technical values, which is not an easy task and the illumination of which is at the core of this article.

2 Elevating Value Sensitive Design

First, this section examines the notion of values and more specifically provides some definitions of the concept of 'value'. This is relevant to the extent that it is human values and their embodiment into technology that is the focus of this paper. The notion of 'values' can be cause for confusion to the extent that values have a mathematical and data science connotation (technical values) as well as used in the sense of morality and ethics (human values). While this paper has its point of departure in the latter notion of values, the core point being argued is that the transformation of human values to mathematical values is not as straightforward as may seem, in turn requiring a balancing act. This section proceeds to investigate the academic discipline of VSD, essentially describing what it is and how it is can be applied.

2.1 Values

In examining a concept for the first time it can be useful to get an initial linguistic meaning or definition. Consequently, 'value' has been described as, 'something (such as a principle or quality) intrinsically valuable or desirable'.[16] It is also described as, '[p]rinciples or standards of behaviour; one's judgement of what is important in life.[17] Values have also been described as what is important to people in their lives, with a focus on ethics and morality.[18]

From the VSD perspective, the notion of what a value is has been described as follows:

> In some sense, we can say that any human activity reflects human values. I drink tea instead of soda. I recently attended a Cezanne exhibit instead of a ball game. I have personal values. We all do. But these are not the type of

[16] Merriam-Webster, *Value*, available at https://www.merriam-webster.com/dictionary/value.

[17] Lexico (Oxford), *Value*, https://www.lexico.com/definition/value.

[18] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 4.

> human values which this volume takes up. Rather, this volume is principally concerned with values that deal with human welfare and justice.[19]

The notion of what a value is, is also dealt with in the realm of legal theory:

> Values are whatever human beings hold to as the underpinning reasons behind more immediate reasons for acting, for approving action, and for preferring certain ways of acting and states of affairs to others. They are as such not necessarily themselves backed by further or ulterior reasons. This we express rather than explain by saying that, for us, something or other is 'good in itself'; whatever is good in itself is, for that person, an ultimate as distinct from a merely instrumental or derivative value. Hence arguments concerning what is of ultimate value cannot proceed by way of demonstration or proof.[20]

Besides examining values in themselves, the relationship between values and technology also encompasses a directional aspect in the manner that entails not merely the embedding of values in technology but also the fact that values and technology exert an influence upon each other. Nissenbaum, in examining the notion of how technology embodies values, highlights the unidirectional manner in which technology exerts an influence over society and which until now has been the predominant focus of scholars. She highlights two trends that exemplify this unidirectional approach: first, there is the situation of computers replacing humans in positions of responsibility, thereby affecting the extent to which society is able to hold humans accountable or responsible (without paying attention to the notion of responsibility as a value in itself); the second is where technological development forces us to re-examine the values themselves, e.g., how should we conceptualize privacy as a value in the light of a new technology. For Nissenbaum, the focus of study should be the opposite, namely, the direction 'from values to technology' and the manner in which values affect technology.[21] She states:

[19] Friedman, Batya, *Introduction*, in Friedman, Batya (ed.) *Human Values and the Design of Computer Technology*, Centre for the Study of Language and Information, 1997, p. 3.

[20] MacCormick, Neil, *H.L.A. Hart*, Stanford: Stanford Univ. Press, 1981, p. 48.

[21] Nissenbaum, Helen, *How Computer Systems Embody Values*, Computer, 2001, available at https://nissenbaum.tech.cornell.edu/papers/embodyvalues.pdf, p. 120.

> Humanists and social scientists can no longer bracket technical details—leaving them to someone else—as they focus on the social effects of technology. Fastidious attention to the before-and-after picture, however richly painted, is not enough. Sometimes a fine-grained understanding of systems—even down to gritty details of architecture, algorithm, code, and possibly the underlying physical characteristics—plays an essential part in describing and explaining the social, ethical, and political dimensions of new information technologies.[22]

It is often the case that when a public uproar erupts over a technological development, it is often the case that the technology has provoked a certain human value held dear by society. This line of argumentation is put forward by Nissenbaum who states that, 'the failure to meet technical criteria [does] not cause the public debate […] it was the controversial ways that these technologies engaged social, ethical and political values that did this.[23] Brownsword and Goodwin refer to certain concepts as 'boundary marking concepts'. These are concepts, they explain, that can be used in discussions about the desirability of certain technologies, and which mark the acceptable border of a technology in relation to morality.[24] In other words, these are concepts that draw the line for what is morally acceptable. Boundary marking concepts have certain distinguishing characteristics. First, they have the goal of being an instrument in determining the boundary of what technology is to be permitted. Second, they are not only concerned with prohibition. Third, a boundary marking concept may have within it a pre-defined notion of what is morally acceptable. For example, the notion that human dignity arises out of a religious belief in itself sets a boundary that is not open for negotiation. Fourth, boundary marking concepts do not exist in a vacuum, but rather reflect the norms of society, which themselves may not remain constant. A certain normative belief system will result in a certain boundary marking concept having a greater importance than another, e.g., those encompassing normative outlooks associated with a religious belief.[25]

[22] Nissenbaum, (n. 21), p. 121.

[23] Ibid., p. 118.

[24] Brownsword, Roger and Goodwin, Morag, *Law and the Technologies of the Twenty-First Century*, Cambridge University Press, 2012, p. 188.

[25] Ibid., p. 190.

2.2 Value Sensitive Design

Turning now to VSD itself, it can be described as a design philosophy that is one of the forebearers of the 'by-design' approach to technology. It is also described as a methodology or approach to the design of information and computer systems that came to the fore in the 1990's.[26] In its most basic form it is described in the following manner:

> Value Sensitive Design seeks to guide the shape of being with technology. It positions researchers, designers, engineers, policy makers, and anyone working at the intersection of technology and society to make insightful investigations into technological innovation in ways that foreground the well-being of human beings and the natural world. Specifically, it provides theory, method, and practice to account for human values in a principled and systematic manner throughout the technical design process.[27]

It is described as, 'a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process'.[28] The main thrust of VSD is to ensure that human values are embedded into the technological design process already from the design stage. The human values that one seeks to embed into the technology are those that have a moral characteristic, e.g. privacy, trust, accountability, honesty, freedom from bias and democracy, to mention but a few.[29] The spectrum of values addressed is relatively wide and may even include values associated with usability, conventions and personal taste.[30]

VSD embodies a number of commitments:

> the relationship between technology and human values is fundamentally interactional; analyses of both direct and indirect stakeholders; distinctions among designer values, values explicitly supported by the project and stake-

[26] Friedman, Batya, *Value Sensitive Design, Berkshire Encyclopedia of Human-Computer Interaction*, 2004, available at https://old.vsdesign.org/publications/pdf/friedman04vsd_encyclopedia.pdf, p. 769.

[27] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 3.

[28] Friedman, Batya, Kahn, Peter and Borning, Alan, Value Sensitive Design: Theory and Methods, UW CSE Technical Report, 2003, available at https://www.researchgate.net/publication/2551270_Value_Sensitive_Design_Theory_and_Methods, p. 1.

[29] Friedman, *Value Sensitive Design*, (n. 26), p. 769.

[30] Ibid., p. 769.

> holder values, individual, group, and societal levels of analysis; integrative and iterative conceptual, technical, and empirical investigations; co-evolution of technology and social structure; and a commitment to progress (not perfection).[31]

Also, important as far as VSD is concerned are the following: it is proactive in nature, it critically assesses human values as it carries them into the design process, it enlarges the scope of human values and it broadens and deepens the methodological approaches, drawing on anthropology, design, human-computer interaction, organizational studies, psychology, philosophy, sociology and software engineering, to mention but a few.[32]

At the core of the VSD design philosophy is a methodology that includes three distinct investigations, namely the investigations of a *conceptual*, *empirical* and *technical* nature, in an integrative and iterative manner. First, regarding the conceptual investigation, it can be said to, 'comprise philosophically informed analyses of the central constructs and issues under investigation […] how does the philosophical literature conceptualize certain values and provide criteria for their assessment and implementation? What values have standing? How should we engage in trade-offs among competing values on the design, implementation and use of information systems […]?[33] The conceptual investigation can be further described by focusing on the types of questions it seeks to address:

> Who are the stakeholders? What is likely to be at stake for people and other nonhuman stakeholders? What theoretical commitments and choice of conceptual framework, if any, are made? If the design team makes a commitment to a particular ethical or cultural framework to support principles reasoning, how would it be articulated and integrated into the design process? What values are likely to be implicated? How will values be framed and characterized? What conceptual models, if any, for operationalizing a given value or values will be employed? How will results from an empirical or technical investigation be integrated into the conceptual framework of

[31] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 4.

[32] Ibid.

[33] Friedman, *Value Sensitive Design*, (n. 26), p. 770. See also Friedman, Batya, Kahn, Peter and Borning, Alan, *Value Sensitive Design: Theory and Methods*, UW CSE Technical Report, 2003, available at https://www.researchgate.net/publication/2551270_Value_Sensitive_Design_Theory_and_Methods, p. 1.

> the project? What value-orientated criteria will be used to judge success of the design?[34]

Additionally, a characteristic of the conceptual investigation is its flexibility, ranging from 'armchair analyses' to more analytical types of investigations. As seen from the above statement, the identification of stakeholders is in focus. Here the conceptual analysis requires the identification of stakeholders, both direct but also indirect, the former category of stakeholder being those that interact directly with a system and the latter being those that are affected by the system, even though they do not use the system.[35] The identification of stakeholders can be important also from the legal point of view. For example, in 2011 the state of Nevada in the USA promulgated a law regulating autonomous vehicles, where rulemaking authority was granted to the Nevada Department of Transportation, which in turn consulted car manufacturers, Google, insurance companies and consumer groups.[36] This example highlights the fact that the stakeholders to a regulatory regime may change and in turn affect the traditional regulatory consultation processes.

One value that is important is that of autonomy. It can be described as, '[…] individuals who are self-determining, who are able to decide, plan, and act in ways that they believe will help them to achieve their goals and promote their values. People value autonomy because it is fundamental to human flourishing and self-development'.[37] However, there are limits to how much autonomy one can provide before autonomy actually starts diminishing. An example is where a product or system is developed for a task, say making presentations. A user will want access to the higher levels of the programme, e.g. how to make slides etc, but not the programme code. The more that the user needs to address at the programme code level, the more autonomy diminishes in that the user will not be able to achieve his or her goals. In this case, autonomy can be described as,

[34] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 32.

[35] Friedman, *Value Sensitive Design*, (n. 26), p. 770. An example provided in the literature is a system used in the health care context, where doctors and nurses would be the direct stakeholders and patients would be indirect stakeholders.

[36] Richards, Neil and Smart, William D., *How Should the Law Think About Robots?*, in Calo, Ryan, Froomkin, Michael A., and Kerr, Ian (eds.), *Robot Law*, Edward Elgar, 2016, p. 12.

[37] Friedman, *Value Sensitive Design*, (n. 3), pp. 17–18.

'when users are given control over the right things at the right time'.[38] Autonomy can therefore be seen in terms of system capability where autonomy can be undermined when the computer system does not provide the user with the necessary technological capability to realize his or her goals.[39] Another value identified in technology is that of bias, a definition being that:

> We say that a computer technology is biased if it systematically and unfairly discriminates against certain individuals or groups of individuals in favour of others. A technology discriminates unfairly if it denies an opportunity or a good, or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate.[40]

The second investigation as part of the VSD methodology is the empirical investigation, which essentially validates and potentially expands the values identified in the conceptual investigation. In other words, while the conceptual investigation may assume a number of relevant values in a specified context, the empirical investigation can refine these values and also confirm the assumptions made in the conceptual investigation. The empirical investigation is required to the extent that the conceptual investigation, 'can only go so far' and that the human context in which the technology operates too needs investigation.[41] Put another way, the reason for applying the empirical investigation is purported to be the limitations connected to applying only a conceptual investigation.[42] Once again, the questions to be asked as part of the empirical investigation are the following:

> How do stakeholders apprehend individual values in the sociotechnical context? How do stakeholders prioritize competing values or otherwise envision resolution of value tensions? Are there differences between espoused practice (what people say) compared with actual practice (what people do)? […] [w]hat are organizations' motivations, methods of training and dissemination, reward structures, and economic incentives?[43]

[38] Friedman, *Value Sensitive Design*, (n. 3), p. 18.

[39] Ibid., p. 18.

[40] Ibid.

[41] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 33.

[42] Ibid.

[43] Ibid., pp. 33–34.

Finally, the technical investigation focuses on the design and performance of the technology itself, where the assumption is that, '[…] technologies in general, and information and computer technologies in particular, provide value "suitabilities" that follow from the properties of the technology […] a given technology is more suitable for certain activities and more readily supports certain values while rendering other activities and values more difficult to realize'.[44] Two additional aspects characterise the technical investigation, namely how technical properties support or hinder human values and secondly the proactive aspect whereby the design properties can be promoted with the intention of promoting the human values identified in the conceptual or empirical investigation.[45] In addressing the technical investigation, the following questions are addressed:

> What features of a technical infrastructure enable, hinder, or even foreclose certain kinds of designs for supporting human activity? How do policies, laws, or regulations create opportunities or constrain options for technological development?[46]

Consequently, VSD design embodies these three investigations that should occur in an integrative and iterative manner. In other words, all these separate investigations in effect influence each other. A description of the VSD methodology reveals that a minimal focus is placed on legal consideration. For example, the questions addressing the technical investigation do consider the role of laws and regulations. However, it is argued that as the awareness of how technology embodies values increases within the legal profession and more importantly within the realm of the legal regulator, so too are the legal requirements increasing that mandate the embedding of legal values (representing human values) in technology. The next section illuminates the EXTREMUM project and the extent to which legal values mandated by law are embedded into the design process of the technological development.

[44] Friedman, *Value Sensitive Design*, (n. 26), p. 770.

[45] Ibid., p. 770.

[46] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 34.

3 Project EXTREMUM

The EXTREMUM project can in simple terms be described as a machine learning initiative whereby useful knowledge is extracted from databases comprising medical data. The knowledge that is sought relates to the adverse effect of certain prescription drugs in order that the adverse effects can be predicted and prevented. The same applies to the detection and predictive treatment of patients in relation to cardiovascular diseases. The ultimate goal of the project is to develop a prototype system that can be used to achieve the above insights from health data and is best explained by an extract from the project web site:

> to develop a novel platform for learning from complex medical data sources with focus on two healthcare application areas: adverse drug event detection and early detection and treatment of cardiovascular diseases [it] will present a new framework for data management and analysis of the integration of data, methods for machine learning as well as ethical issues related to predictive models. The fundamental breakthrough of this project is to establish a novel knowledge management and discovery framework for medical data sources. The outcome will be a set of methods and tools for integrating complex medical data sources, a set of predictive models for learning from these sources with emphasis on interpretability and explanatory features, and simultaneously focusing on maintaining ethical integrity in the underlying decision mechanisms that rule the machine learning.[47]

From the above text it becomes apparent that the tools used to extract knowledge from the medical data are complex machine learning algorithms and models. What the VSD philosophy and methodology shows us is that these algorithms and models are not value-neutral to the extent that they include social, ethical and moral values – human values. The next consideration entails which values should be integrated or embedded into these machine learning algorithms and models. The point of departure is naturally that there is a wide spectrum of values that could potentially be relevant, depending on which stakeholders you address as well as what context one is considering. Which values should therefore be included for consideration? In order to simplify matters, and from the legal perspective, the answer to this question is rather obvious – it is the human values as mandated and promoted in the formally promulgated

[47] Project EXTREMUM, https://www.digitalfutures.kth.se/research/collaborative-projects/extremum/.

laws that regulate a relevant context. Once again, not all laws can be consulted and not all human values addressed in these laws can be taken into account. For example, it may not only be formal legal instruments that promote values but other regulatory instruments that reside under the banner 'soft law' may also be relevant.[48] However, employing the VSD technical investigation, and also based on legal experience, one soon recognizes which legal frameworks are more relevant in the given context. For example, a project working with personal data in the form of health data naturally calls into consideration the General Data Protection Regulation (GDPR) and consequently the values this legal instrument promotes.[49] Even within the GDPR a wide range of human values may exist. The legal values addressed in the next section have been selected in order to illuminate the main goal of this paper and have therefore been chosen not only because of their relevance but also because of the pedagogical value that a discussion of these values promotes.

3.1 Identified Legal Values

Three human values have been identified as being relevant to the extent that they are mandated by the GDPR. These three values are *explainability*, *privacy* and *accuracy*. An in-depth analysis of these three concepts or values remains beyond the boundaries of this paper and their choice is merely illustrative of the fact that human values are mandated by legal instruments when designing technology. There are many human values promoted by the GDPR, e.g., autonomy, personal integrity and dignity, to name but a few. The three values of explainability, privacy and accuracy have been chosen as they have one thing in common – that is, echoing Friedman above, they promote human welfare and justice. This said, a very brief explanation of the above human values follows.

A central principle of the GDPR is that the data subject be granted information concerning the processing of personal data. More specifically, Article 13(2)(f), 14(2)(g) and 15(1)(h) regulate the provision of in-

[48] Here an example of a 'soft law' code is the *Ethics Guidelines for Trustworthy AI* by the European Commission High-Level Expert Working Group on Artificial Intelligence, available at https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

[49] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016.

formation in relation to Articles 22 concerning automated decisions. In addition, Recital 63 is relevant to the extent that the data subject should have a right of access to, 'the logic involved in any automatic personal data processing and, at least when based on profiling'. In relation to Article 22, a reference to explainability is found in Recital 71, which states that the data subject has, 'the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision'. Whether the explanation is formed as a right is beyond the bounds of this paper.[50] What is noteworthy for the purposes of this paper is that explainability exists as a value that is promoted by the GDPR. And the justification for enshrining explainability is illuminated by Bygrave in his reference to opaque machine learning decisional systems and the human interest in 'cognitive sovereignty', stating that, '[t]he interest is foundational to the normative justification for requiring explicability of machine processes'.[51] It is argued that there is still considerable discussion concerning what explainability actually entails and many questions arise: what is meant by this explainability? Explainability for which stakeholder? Is it explainability for data scientists or data subjects? And is explainability even attainable in the era of deep neural networks, the inner workings of which go beyond the cognitive ability of human beings?

The questions are complex and the answers elusive. However, the intention of this paper's emersion in the notion of values such as explainabilty is echoed by Brkan and Bonnet:

> The GDPR is thus becoming increasingly important also for XAI researchers and algorithm developers, since the introduction of the legal requirement for understanding the logic and hence explanation of algorithmic de-

[50] For an in-depth discussion surrounding explainability as a right, see Goodman B and Flaxman S, *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"*, ICML Workshop on Human Interpretability in Machine Learning, arXiv:1606.08813, AI Magazine, Vol 38, No 3, 2017, Wachter S, Mittelstadt B, and Floridi L, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, in International Data Privacy Law, Volume 7, Issue 2, May 2017, pp. 76–99 and Selbst, A and Powles J, Meaningful Information and the Right to Explanation, in International Data Privacy Law, Vol. 7, No. 4, 2017, pp. 233–242.

[51] Bygrave, *Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions*, (n. 15), p. 8.

> cisions entails also the requirement to guarantee the practical feasibility of such explanations from a computer science perspective.[52]

The value of privacy is promoted by the GDPR in its entirety. However, one of the main Articles in which it finds expression is Article 25, entitled 'Data protection by design and by default'. The connection between Article 25 and the notion of privacy is indisputable. In the words of Bygrave, '[a]rticle 25 springs out of a policy discourse that commonly goes under the nomenclature 'Privacy by Design' ('PbD').[53] Also, in 2010 the 32nd International Conference of Data Protection and Privacy Commissioners passed a resolution to the effect that Privacy by Design was a fundamental part of fundamental privacy protection.[54] Ann Cavoukian is credited with coining the notion 'Privacy by Design', which is briefly described as, '… an approach to protecting privacy by embedding it into the design specifications of technologies, business practices, and physical infrastructures. That means building in privacy up front – right into the design specifications and architecture of new systems and processes'.[55] The main rationale to data protection by design is that privacy-related interests receive serious consideration throughout the entire lifecycle of information systems development and not just at the end.[56]

Article 25(1) states:

> Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary

[52] Brkan, M., and Bonnet, G., *Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas*, European Journal of Risk Regulation, Vol. 11, Issue 1, March 2020, pp. 18–50, p. 19.

[53] Bygrave, Lee A., *Data protection by design and by default*, in Kuner Christopher, Bygrave Lee A. and Docksey Christopher (eds.), *The EU General Data Protection Regulation (GDPR)*, Oxford University Press, UK, 2020, p. 571.

[54] Ibid., p. 571.

[55] Information and Privacy Commissioner of Ontario, *Introduction to PbD*, available at https://www.ipc.on.ca/english/privacy/introduction-to-pbd/ (last accessed on 2016-03-24).

[56] Bygrave, *Data protection by design and by default*, (n. 53), p. 571.

> safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

For the purposes of this paper, it is sufficient to note that privacy is an inherent value protected and promoted by the GDPR and that it finds expression throughout the GDPR but also very concretely in Article 25 GDPR.

The GDPR incorporates a number of data protection principles in Article 5, and in Article 5(1)(d) accuracy is mentioned as a principle. The rationale for incorporating accuracy as a principle is brought to the fore and expanded upon by the Article 29 Data Protection Working ('Working Party') Party.[57] The Working party specifically mentions accuracy in relation to profiling and that it should be taken into account during the collection of data, the analysis of data, the building of a profile and the application of a profile.[58] The rationale for the need for accuracy is provided by the Working Group:

> If the data used in an automated decision-making or profiling process is inaccurate, any resultant decision or profile will be flawed. Decisions may be made on the basis of outdated data or the incorrect interpretation of external data. Inaccuracies may lead to inappropriate predictions or statements about, for example, someone's health, credit or insurance risk.[59]

Finally, it is argued that profiling may include an element of prediction, which in turn can result in inaccuracies if the underlying data is incorrect.[60] Here is can be argued that accuracy is necessary in order that an algorithm, creating a profile of an individual, paints as true a picture as possible of that individual. Accuracy can therefore be seen as a component necessary to gauge a person's reputation.[61] In this sense it is a human value worth preserving and promoting.

[57] Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, Adopted on 3 October 2017 available at wp251rev_01_en_A754F3E1-FB46-9E76-C0A919864E4B6641_49826.pdf.

[58] Ibid., p. 12.

[59] Ibid.

[60] Ibid., p. 17.

[61] For a discussion on reputation, see Greenstein, Stanley, *Our Humanity Exposed: Predictive Modelling in a Legal Context*, Dissertation, Stockholm University, 2017, available at http://www.diva-portal.org/smash/record.jsf?dswid=5270&pid=diva2%3A1088890.

The next section entails a closer examination of how the values of explainability, privacy and accuracy to some extent compete with each other but also complement each other.

3.2 The Trade-Off in Transforming Legal Values to Technical Values

This section draws on experiences from the EXTREMUM project in order to illustrate the challenges associated with transforming human values or legal values expressed in the natural language form into technical values represented by machines. The process of transforming human values into technical values is depicted in Figure 1 below. It is argued that the three legal values mentioned above, namely, explainability, privacy and accuracy, can be categorized into three separate techniques: the first is the learning of a machine learning classifier that can accurately diagnose patients based on historical patient data, the second is preserving the privacy of patients' data, and the third is providing explanations to diagnoses made by the developed machine learning models.[62]

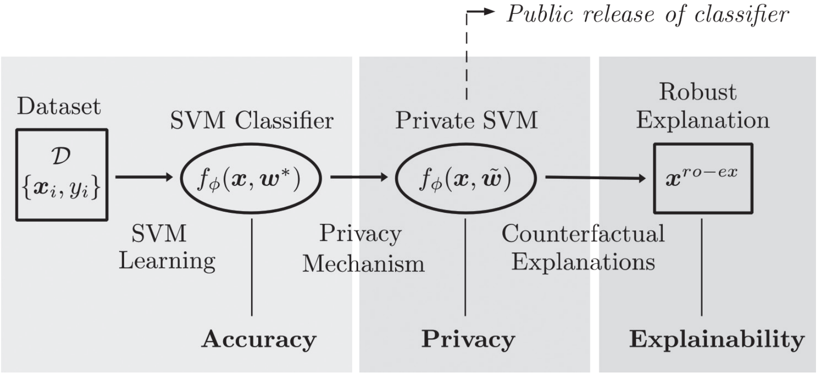

Public release of classifier
Dataset
$\mathcal{D}$
$\{\boldsymbol{x}_i, y_i\}$
SVM Classifier
$f_\phi(\boldsymbol{x}, \boldsymbol{w}^*)$
Private SVM
$f_\phi(\boldsymbol{x}, \tilde{\boldsymbol{w}})$
Robust Explanation
$\boldsymbol{x}^{ro-ex}$
SVM Learning
Privacy Mechanism
Counterfactual Explanations
Accuracy
Privacy
Explainability

Figure 1 is a representation of the machine learning techniques performed in order to gain the insights from medical data in accordance with the goals of the EXTREMUM project.

In Figure 1, a dataset containing sensitive data with many patients' historical health records and respective diagnosis, represented by 'D', is se-

[62] Mochaourab, R, Sinha, S., Greenstein, S. and Papapetrou, P., *Robust Counterfactual Explanations for Privacy-Preserving SVM*, International Conference on Machine Learning (ICML 2021), Workshop on Socially Responsible Machine Learning, Jul. 2021.

lected. It is assumed that the dataset is securely stored within the confines of the hospital's technical infrastructure without public access to its entries. Within the secure confines of the hospital, the dataset is employed to train a machine learning classifier, in this case a Support Vector Machine (SVM),[63] that would predict the diagnosis of future patients based on the available historical data. The objective of SVM learning is to achieve the highest possible prediction accuracy and accordingly perform correct diagnosis for most future patients, i.e. as many patients as is technically possible. Hence, the value of accuracy is the prime goal.

The functionality of the SVM classifier depends on a set of parameters which are determined using the patients' health records in the dataset. Hence, any public accessibility to the trained SVM classifier parameters may lead to privacy breaches if an adversary manages to reconstruct the patients' dataset using the classifier parameters. Therefore, we need to ensure that the privacy of the persons in the dataset is preserved before publicly releasing the classifier. The privacy mechanism used here guarantees differential privacy, which is a privacy mechanism that incorporates a tuneable degree of uncertainty about the actual presence of any entry in the dataset (also referred to as 'noise').[64] This uncertainty is achieved through random perturbation of the SVM classifier parameters.[65] Consequently, the private version of the SVM classifier can be made publicly available with potential utilization in various contexts, e.g., in many different hospitals.

The benefit in guaranteeing privacy comes at the cost of reduced classifier accuracy. In other words, these two technical values are mutually exclusive to the extent that increasing one of them decreases the other. Figure 2(a) below illustrates the differences between the optimal SVM classifier and its private version. Firstly, two categories of patients are represented. The category of 'healthy patients' is represented by the circles and the category 'unhealthy patients' is represented by the plus signs. The two axes in the figure represent two features of the data, i.e., two attributes of the patients' health records. Examples of features can be smoking,

[63] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference and prediction*. Springer Series in Statistics. Springer-Verlag New York, 2nd edition, 2009.

[64] Dwork, C. and Roth, A. *The algorithmic foundations of differential privacy*. Found. Trends Theor. Comput. Sci., 9(3–4):211–407, August 2014. ISSN 1551-305X.

[65] Pertubation can be described as the adding of noise to data in order to enhance confidentiality.

amount of exercise, genetic indicators or previous health issues, to name a few. The optimal SVM decision boundary separates the two of patients classes with a straight line which maximizes the widths of the margins. Accordingly, the optimal SVM decision boundary is robust to any small changes in the data since it is furthest away from both sets of points (i.e. both circles and plus signs). On the other hand, the private SVM boundary (indicated with the broken line), which is a randomly perturbed version of the optimal SVM, is clearly less robust than the optimal SVM boundary (its distance to the data sets has diminished and is therefore not as robust). This means that its accuracy in classifying future patient cases may on average be lower than that of the optimal SVM while on the other hand it enhances privacy. In this way, the striving after privacy in technical terms comes at the expense of accuracy.

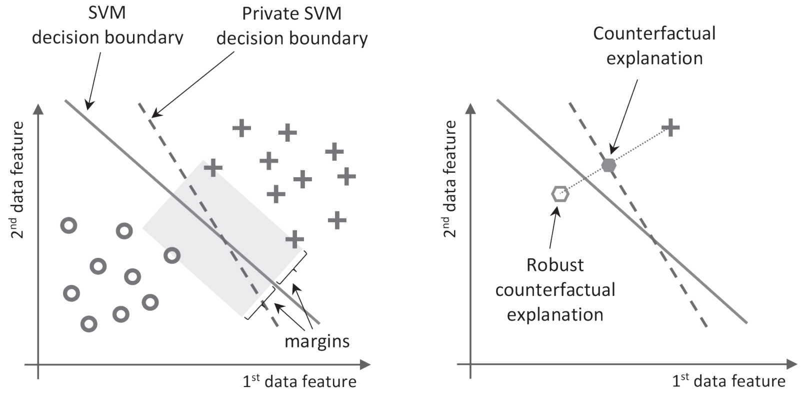


Figure 2(a)

Figure 2(b)

Figure 2 (a) illustrates the machine learning classifiers and the differences between the optimal and private SVM. Figure 2 (b) shows a counterfactual explanation for a single data point as well as its robust version.

Figure 1, also proceeds to consider the explainability of classifications made by the private SVM (privacy preserving SVM). This can be explained by a hypothetical example depicting the patient and medical practitioner context. On informing a patient about the diagnosis provided by the classifier, it becomes possible to quantify the least necessary changes to the patient's data that would lead to another diagnosis. For example, this could entail informing a patient of what is required for him or her to move from the category of ill patients to the category of healthy patients. This can be related to Figure 2(b), where the changes lead to a given patient's data point on one side of the SVM decision boundary moving to the opposite side of the SVM decision boundary. Subsequently, the patient perceives a contrastive example of related data which helps in explaining the diagnosis. Such types of explanations are called counterfactual explanations and are especially useful when the changes in the data are actionable, i.e., the patient is able to perform certain tasks to change the outcome of the diagnosis.[66]

Providing valid counterfactual explanations for the privacy preserving SVM classifier is a challenging task due to the introduced perturbations to the optimal classifier parameters. Observe that the optimal SVM classifier is unknown since only the private SVM is publicly released. The necessary changes to the patient's data that lead to a different diagnosis according to the private SVM classifier may still give the same original diagnosis according to the optimal SVM classifier, as is shown in Figure 2(b). Addressing this issue requires studying robust counterfactual explanations that consider the extent of perturbations required by the privacy mechanism.[67] Generating larger perturbations to achieve larger levels of differential privacy would essentially require larger changes in the patient's data for counterfactual explanations. This is illustrated in Figure 2(b) where the robust counterfactual explanation is further away on the other side of the boundary to make sure that it is correctly classified according to the optimal SVM decision boundary. In other words, these larger changes would guarantee a desired level of confidence that we predict a different diagnosis using the unknown optimal SVM classifier. Hence, as a summary, guaranteeing a desired level of differential privacy diminishes the classifier accuracy and consequently increases the required changes in counterfactual explanations to meet a certain level of confidence in validity of the explanations.

[66] Wachter, Sandra, Mittelstadt, Brent C. R., *Counterfactual explanations without opening the black box: Automated decisions and the GDPR*. Harvard Journal of Law & Technology, 31(2), 2018.

[67] In this regard reference is made to footnote 62 above.

4 Conclusions

This paper began with the argument that whether we like it or not, the technology we create has embedded within it human values, more specifically social, ethical and moral values. There may be varying interpretations as to what a value is and where the boundaries lie as far as values are concerned, but for the sake of simplicity, a value can be said to be a good in itself. The paper then proceeded to illuminate the philosophy of VSD, which places a large emphasis on identifying the human values associated with technological development. However, from the purely legal perspective, the human values inherent in regulatory instruments provide a natural point of departure for a discussion of values.

Using knowledge from research within the EXTREMUM project, the main argument put forward in this article is that having identified the human values relevant in relation to a particular technology and social context, it may not be a straightforward issue of transposing these into the language of data science. Challenges include the fact that the values expressed in laws have not undergone a balancing process. In other words, the GDPR refers to explainability, privacy and accuracy but it does not consider the difficulties in embedding these values into the technology. However, these difficulties become apparent when the process of the translation of the values from natural language to the language of data science begins. It soon becomes apparent that in technical terms these human values are mutually exclusive – promoting one will invariably occur at the expense of another – which in turn leads to the next problem of having to balance these competing human values against one another as they are transformed into their mathematical equivalents. This is depicted in the EXTREMUM project where the human values of privacy and accuracy are pitted against each other as they are transformed into their mathematical and statistical equivalents, or put another way, into the rules of data science. Adding to the picture is the value of explainability which adds a layer of complexity.

Much attention is given to the fact that cross-disciplinary work is required to address the challenges of modern technological development. However, this is easier said than done, as depicted by the findings highlighted in this paper. One issue that comes to the surface is extremely important. The data scientist's reflex is to focus on the value of accuracy, which is nothing strange as his or her main focus is to develop technology that works in a manner that is as accurate as possible. However, professionals from other disciplines such as the law, bring other insights to the table regarding human values and legal demands, which in turn will focus attention on other values in addition to accuracy, privacy being a prime example. The fact that values such as privacy are mandated by the law essentially create a dilemma for the data scientist. Naturally, the technology should be as accurate as possible, but adhering to the law may mean that part of the accuracy must be sacrificed for the sake of privacy – not because we want to, but because we have to – if we want to follow the law, that is. The follow-up question is extremely interesting, namely, if we then are mandated by law to insert privacy into technology, how much privacy is enough privacy according to the law? This in turn raises additional questions, e.g. how to quantify privacy and what level of privacy is demanded by the law? These are questions that remain outside the boundaries of this paper but that will hopefully be addresses in future works and fora.

It is argued that the above experiment brings worthwhile insights from various perspectives: it can be worthwhile from the legislative technique's perspective, i.e. in relation to how we can better produce laws and other sources of law; it is a valuable insight for legal practitioners called upon to ensure that human values and more specifically legal values are embedded into the technology we create and finally it creates awareness surrounding the insight that the technology we create reflects the human values we embrace.

Katja de Vries

A Researcher's Guide for Using Personal Data and Non-Personal Data Surrogates: Synthetic Data and Data of Deceased People

1 Introduction

In 2018, I was working as a postdoc at the IT University (ITU) in Copenhagen. Besides teaching and research, I also sometimes acted as a mediator between the Data Protection Officer[1] (DPO) and researchers, trying to match the legal requirements following from data protection law with the practical concerns that researchers face when handling data in the, often, messy realities of doing research. In early 2020, in a Preliminary Opinion on data protection and scientific research,[2] the European Data Protection Supervisor (EDPS) wrote:

> Data protection rules aim to ensure safety and transparency while minimising interference with ethical research that aim at generalisable knowledge and societal good. The GDPR serves in part to ensure accountability for such practices. There is no evidence that the GDPR itself hampers genuine scientific research.

[1] I would like to thank Rasmus Balle and Lilian Baunbæk (that is, the Data Protection Officer and Chief Information Security Officer at ITU during that period of time) for the pleasant and insightful collaboration.

[2] European Data Protection Supervisor (EDPS), *Preliminary Opinion on data protection and scientific research*, 6 January 2020.

Does the GDPR indeed not hamper research? Talking to researchers about the GDPR has given me first-hand experience of what it must feel like to be a dentist handling anxious patients: upon entering a room, I could feel the waves of GDPR-anxiety flowing towards me. While it is true that the General Data Protection Regulation 2016/679[3] (GDPR) is much more research-friendly than many researchers might think, it does require that researchers do some substantive and pre-emptive thinking about the data they plan to process in their research. They have to ask themselves fundamental questions such as: Do I really need this data? Can I do my research in a less data-intensive way? Can I use pseudonymised or anonymised data instead? How long do I need to keep this data? What is the exact purpose that the data fulfil in my research? How can I notify data subjects about the data processing? Are the technological and organisational measures that I have taken to keep the data secure appropriate for the state-of-the-art; the potential risks to the rights and freedoms of the people whose data are processed; and the nature, scope, context and purposes of processing? If such questions are taken seriously, they require some real thinking and balancing of interests, and are not exhausted by simply ticking a box. No wonder that researchers, if they have the choice, prefer using data that fall outside the scope of the GDPR and that allow them to skip the GDPR-based soul searching about their research. This contribution reaches out a hand to researchers, especially those processing data for the creation of AI-models; moreover, it takes them on a quick tour through their options for finding data to fuel their research. In section 2, I revisit the question of whether data protection is stifling AI-research. Then in section 3, I look at initiatives that the EU legislator has recently proposed to make the lives of researchers easier, while staying within the boundaries of the GDPR, notably through the concepts of ‘data intermediation services’ and ‘data altruism’ in the proposed Data Governance Act. Thereafter, in section 4 and 5, I look at two possible surrogates for personal data, which allow researchers to escape from the scope of the GDPR: data of deceased people and synthetic data. Finally, in section 6, I present some concise conclusions and pointers for the researcher suffering from GDPR-anxiety.

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and the repealing Directive 95/46/EC, OJ L 119, 4.5.2016, 1–88.

2 Is data protection stifling AI research and innovation?

All over the world, countries are fighting to stay ahead in the Artificial Intelligence (AI) race[4] and to create a climate that stimulates AI, in terms of research and innovation as well as its adoption and commercialisation. In order to create and use AI, it is indispensable to have data, preferably of good quality and in abundant amounts. For example, to create an AI or machine learning (ML) model that can separate different types of tumours on brain scans, healthy lungs from those affected by Covid or that can recognise the author of an anonymous text, one needs to have enough training data that a model can learn from. Thus, access to data is the fuel for keeping up in the AI race.

Information technology allows data to move around quickly and seemingly effortlessly. Data, however, never move in a legal vacuum. Legal regulation mandates if data are allowed to move freely or not, and under what conditions. Legal fields that decide if data movement is permitted, obligatory, conditional or prohibited include data protection, intellectual property, the right of access to public documents based on the principle of open government, the right to re-use of publicly funded information as open data, and research ethics regulations.

Research and innovation are areas that are prioritised and fostered within the EU. While access to data can be limited by rights of others, such as intellectual property or data protection rights, the EU legislator often reserves a special regime for activities that fall in the field of research and/or innovation. To illustrate this, I will name four (proposed) legal EU instruments that all give a privileged status to data used for research purposes: the Text- and Data mining exception in the Copyright Directive, platform access in the proposed Digital Services Act, the exclusion of research from the scope of the proposed AI Act, and the research exception in the GDPR.

The first example of a research exception relates to data that are protected as works by copyright (for example, tweets, pictures or drawings that contain at least a minimal trace of authorship) or as databases by the

[4] Daniel Castro & Michael McLaughlin, Who Is Winning the AI Race: China, the EU, or the United States? – 2021 Update. Information Technology & Innovation Foundation (ITIF), at: https://itif.org/publications/2021/01/25/who-winning-ai-race-china-eu-or-united-states-2021-update (published online 25 January 2021).

sui generis database-right in Database Directive 96/9.[5] Normally, such works or databases cannot be used without permission from the holder of the intellectual property right. However, Article 3(1) of the Copyright Directive 2019/790[6] contains an exception for Text- and Data mining (TDM) 'for the purposes of scientific research', where scientific research is understood as non-commercial research.[7] This means that non-commercial researchers can train AI models on protected works and databases without needing permission from the rightsholder.

The second example is the right to platform data access which can be found in Article 31(2) and (4) of the proposed Digital Services Act,[8] and gives 'vetted researchers'[9] access to data 'for the sole purpose of conducting research that contributes to the identification and understanding of systemic risks'. While this would be helpful to certain researchers, for example, those creating AI models that capture the spread of disinformation of platforms, such as Facebook or Twitter, several commentators have criticised the narrow scope of this exception and proposed that it should be broadened, both in terms of types of researchers and research.[10]

The third example is the exclusion of research from the scope of the proposed AI Act.[11] The AI Act aims to regulate AI systems that impact on

[5] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, 20–28.

[6] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and the amending Directives 96/9/EC and 2001/29/EC, OJ L 130, 17.5.2019, 92–125.

[7] Recital 12 of the Copyright Directive 2019/790. This narrow interpretation of "scientific research" is not uncontroversial. See e.g. Rossana Ducato and Alain Strowel, Ensuring text and data mining: Remaining issues with the EU copyright exceptions and possible ways out, 43 European Intellectual Property Review, 5, 2021, 322–337.

[8] Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and the amending Directive 2000/31/EC, COM/2020/825 final, 15 December 2020.

[9] According to Article 31(4) DSA, "vetted" means that researchers are 'affiliated with academic institutions, be independent from commercial interests, have proven records of expertise in the fields related to the risks investigated or related research methodologies, and shall commit and be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request'.

[10] Paddy Leersen, Platform research access in Article 31 of the Digital Services Act. Sword without a shield? Verfassungsblog: On matters constitutional, at: https://verfassungsblog.de/power-dsa-dma-14/ (published online 7 September 2021).

[11] European Commission, Proposal for a Regulation of the European Parliament and

society and citizens, potentially in a negative way, which means that the AI Act only concerns AI systems *in practice*, that is, those that are placed on the market, into service or used (Article 1(a) and (b)). Specifically, the adjustments[12] proposed by the Council in Article 2(6) and (7) underline the special status of AI research and development:

> Article 2(6). This Regulation shall not apply to AI systems, including their output, specifically developed and put into service for the sole purpose of scientific research and development.
>
> Article 2(7). This Regulation shall not affect any research and development activity regarding AI systems in so far as such activity does not lead to or entail placing an AI.

The new Recital 12a fleshes this point out even further by saying that the AI Act should not apply to AI systems, which are used for 'the sole purpose of research and development' in order to ensure that the Act 'does not otherwise affect scientific research and development activity on AI systems' but 'that any other AI system that may be used for the conduct of any research and development activity should remain subject to the provisions'. Thus, the Council's adjustments show that the European legislator feels the need to stress that the regulation of certain risky AI practices should not hinder AI systems with the sole purpose of research and development.

In all the aforementioned research exceptions, the research that is protected is somehow limited. The TDM-exception on copyright and database rights is limited to non-commercial science, the platform access exception only can be used by 'vetted' academics that study 'systemic risks' related to the platform data, and the AI Act only excludes AI systems that are completely disconnected from practice, whose sole purpose is research and development. In comparison, the 'scientific research' ex-

of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, COM(2021) 206 final, Brussels, 21 April 2021.

[12] Council of the European Union, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – Presidency compromise text, 2021/0106(COD), Brussels 29 November 2021.

ception in Art. 89 of the GDPR is much wider in scope, 'including for example technological development and demonstration, fundamental research, applied research and privately funded research' (Recital 159). The question arises: how broad is the scope of the scientific research exception exactly? This question has become even more important since the GDPR has come into force. In contrast to its predecessor, Data Protection Directive 95/46[13], the GDPR includes a scientific research exception for the processing of *sensitive* personal data (Article 9(2)j GDPR), such as data relating to health or race. Do practices that mix commercial exploitation with research also qualify as 'scientific research', in the meaning of the GDPR? Scandals like the NHS/DeepMind deal from 2016[14] raise the question as to what kind of research should benefit from the GDPR exception. Should companies that can mix treatment, research and commercial development of health-related AI, such as *Google Health*, fall under the privileged scientific research regime of the GDPR? In a recent preliminary Opinion, the European Data Protection Supervisor (EDPS) does not exclude commercial research as such but argues that the scope of the exception in the GDPR should be limited to research that is 'set up in accordance with relevant sector-related methodological and ethical standards', which includes 'the notion of informed consent, accountability and oversight' and that 'is carried out with the aim of growing society's collective knowledge and wellbeing, as opposed to serving primarily one or several private interests'.[15] Even if one follows this narrower reading of the scope of the scientific research, the scope is still very broad in comparison to the other research exceptions discussed earlier in this section. Certain types of research performed by a commercial company like *Google Health* might very well fall within the EDPS definition of 'scientific research'. However, research that is only commercial and does not follow relevant ethical or medical standards, will fall outside.

[13] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23 November 1995, 31–50.

[14] Julia Powels, Why are we giving away our most sensitive health data to Google? The Guardian, at: https://www.theguardian.com/commentisfree/2017/jul/05/sensitive-health-information-deepmind-google (published 5 July 2017); Julia Powles & Hal Hodson, Google DeepMind and healthcare in an age of algorithms. 7 Health and Technology, Issue 4, 2017, 351–367.

[15] EDPS, *A Preliminary Opinion* (n. 2), 11–12.

The broad understanding of 'scientific research' in the GDPR is all the more important because the scope of the GDPR itself is very large: personal data, that is 'any information relating to an identified or identifiable natural person' (Article 4(1) GDPR[16]), is a category of data that is extremely broad, as underlined, for example, by Purtova.[17] Even data that, at first sight, do not seem to be personal can still qualify as personal data if there is a reasonably likely potential that the data could be retraced to a living individual: I might not *directly* recognise a person from an IP address, a movement pattern or a brain scan, but with some additional research and by combining data from other data, I would be able to connect the dots. This implies that an enormous amount of research falls within the scope of the General Data Protection Regulation (GDPR) 2016/679.[18] However, in order to establish if the GDPR creates a hindrance for AI research, it is not enough to look at the scope of the exception. It is the content of the exception in Article 89 GDPR that is most crucial. The research exception entails that researchers fall under a lighter regime of data protection and have to comply with fewer requirements. Especially, the informational rights of data subjects (right to have data rectified, to access the data, right to object to the processing or to restrict it, right to erasure, etc.) are much more limited or sometimes even non-existent when it can be shown that the exercise of such rights interferes with the research. Yet, as mentioned in the introduction, the data protection requirements that have to be complied with within this privileged regime still compel substantive, pre-emptive thinking about matters like data minimisation and purpose limitation and this often instils a sense of GDPR anxiety in researchers.[19] Does this mean, notwithstanding that the research exception in the GDPR is much more generous in scope than any of the other exceptions mentioned above, that the EU has created a hurdle for itself, and that innovation might be stifled by data protection requirements? While some argue that this is the

[16] GDPR 2016/679 (n. 3).

[17] Nadezhda Purtova, The law of everything. Broad concept of personal data and future of EU data protection law, 10 Law, Innovation and Technology, 2018, 40–81.

[18] GDPR 2016/679 (n. 3).

[19] Katharina Ó Cathaoir, Hrefna Dögg Gunnarsdóttir & Mette Hartlev, The journey of research data: Accessing Nordic health data for the purposes of developing an algorithm, *Medical Law International*, 2021, 1–23.

case,[20] others consider GDPR anxiety to be a result of misunderstandings and hence a transitional matter – the requirements are not unreasonably burdensome, and it simply takes a while to get used to GDPR compliance. Here, an analogy with environmental law[21] could be made, as the concern that the GDPR will hinder innovation resembles the ones raised during the 1970s and 80s about how environmental laws could turn out to be extremely detrimental for businesses and international competitiveness. These concerns about environmental regulation have been refuted,[22] and the benefits of regulating an industrial wild west outweigh the costs: if left unregulated, the behaviour of industry is likely to result in an environmental tragedy of the commons. The same arguments are raised to defend data protection and the regulation of AI.[23] While the analogy between environmental and data protection works well in many respects, there are at least two important differences between the latter and the former. Firstly, the number of actors that is affected by data protection is much larger. Data protection does not only affect big companies like *Amazon*, *Google* or *Meta*; it also affects any individuals or clubs who publish information about other people on their private website, the local

[20] Tal Zarsky, Incompatible: the GDPR in the age of big data, 47 Seton Hall L. Rev., 2016, 995–1020.

[21] This is an analogy that has been made quite frequently in academic literature, although the main point of comparison tends to be that the regulation of data protection and environmental protection use similar regulatory tools (risk impact assessments, transparency requirements, etc.) and that a lack of regulation leads to a tragedy of the commons in both fields. See e.g.: Magdalena Słok-Wódkowska & Joanna Mazur, Regulating the digital environment: what can data protection law learn from environmental law?, 19 Review of International, European and Comparative Law, 2021, 13–43; Mary Julia Emanuel, Evaluation of US and EU Data Protection Policies Based on Principles Drawn from US Environmental Law in: D. Svantesson & D. Kloza (eds), Trans-Atlantic Data Privacy Relations as a Challenge for Democracy, Cambridge 2017, 407–427; A. Michael Froomkin, Regulating Mass Surveillance as Privacy Pollution: Learning from Environmental Impact Statements, University of Illinois Law Review, Issue 5, 2015, 1713–1790; Dennis D. Hirsch, Protecting the Inner Environment: What Privacy Regulation Can Learn from Environmental Law, 41 Georgia Law Review, Issue 1, 2016, 1–63.

[22] Frank S Arnold, Anne S Forrest & Stephen R Dujack, *Environmental Protection: Is it Bad for the Economy?*: Environmental Law Institute 1998.

[23] See, for example, the extensive report produced by the Dutch Scientific Council for Government Policy: Corien Prins, Haroon Sheikh, Erik Schrijvers, Eline De Jong, Monique Steijns & Mark Bovens. *Mission AI. The New System Technology. Summary report*, Wetenschappelijke Raad voor het Regeringsbeleid (Scientific Council for Government Policy), The Hague, Netherlands, 2021.

supermarket that has a customer loyalty programme, the municipality dealing with citizens, researchers doing research involving personal data, etc. The list is, in principle, endless, and it is currently difficult to find anyone inside the EU who has not heard or been affected by the GDPR. Even if I only focus on personal data processing in *AI research*, which is only a tiny domain within the much larger material scope of the GDPR, the impact of the GDPR is enormous in comparison to environmental regulation of industry. This brings me to a second difference, which I have already discussed above: namely that even under the lighter regime of the research exception (Article 89 GDPR), the GDPR still requires individual researchers to do some fundamental and substantive thinking about the data needed for their research. Compared to a data protection impact assessment (looking at the risks to individual *rights*),[24] an environmental impact assessment (looking at the risks to the *environment*) has a more tangible and quantifiable object Loss in biodiversity or increase in CO_2 can be quantified. The numbers might be up for debate but at least some quantification is possible. Quantifying the risk to a right is more difficult: how to put a number to how much privacy is lost, how much more cautious citizens become due to an increased chilling effect, or how much of the rule of law evaporates? Researchers processing personal data will often have to do some balancing of interests in light of a deep understanding of their research and its impact. A DPO can *assist* a researcher in asking the right GDPR questions, such as 'Is the public interest pursued important enough to legitimise the negative effects for affected individuals?' However, the answer to these questions can only be found by *combining* detailed knowledge about the research set-up and purposes with an understanding of data protection law. Of course, one should not exaggerate the burden of data protection compliance. Thinking seriously about research data (how you use them, how long you need to keep them, why you need them, how to keep them secure, how you notify affected data subjects, etc.) should, in principle, not be an insurmountable burden. Yet, in the busy day-to-day life of many researchers, the discovery that GDPR-compliance requires more than the thoughtless ticking of a few boxes, combined with the fear of hefty administrative fines, in the case of non-compliance (Article 83 GDPR), can cause the

[24] Raphael Gellert, *The Risk-based Approach to Data Protection*, Oxford University Press (2020).

aforementioned GDPR anxiety. Moreover, some researchers[25] might argue that notably the principle of *purpose specification* in Article 5(1)(b) GDPR, which requires personal data to be collected for a 'specific, explicit and legitimate' purpose, and *data minimisation* in Article 5(1)(c), which requires that data be 'limited to what is necessary in relation to the purposes for which they are processed', are at odds with the flexible attitude underlying much big data research that boils down to 'Let's gather as much data as I can, and then just try out some things – I'll find out by trial and error what generates interesting results'.

3 How the EU legislator wants to make the life of researchers easier: Data intermediation services and data altruism

Is GDPR anxiety just another name for a sloppy research attitude, entailing the lack of proper hypotheses and research plans, and a too limited understanding of the opportunities offered to researchers in the GDPR? It might be – in some cases, at least – but the EU legislator clearly feels the need to help researchers by making data processing within the boundaries of the GDPR easier. One of the proposals that could help to realise these ambitions is the proposed Data Governance Act[26] (DGA), presented by the Commission in November 2020. On 31 November 2021, the European Parliament and Council reached an agreement and presented a provisional final version.[27] In Recital 5 of this latest version of the proposed DGA, it states in relevant part:

[25] Zarsky (n. 20).

[26] European Commission, Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act), COM(2020) 767 final, Brussels, 25 November 2020. It should be noted that the DGA not only tries to facilitate the sharing of personal data, but also data which are protected by intellectual property rights. In this contribution, I only focus on the sharing of personal data in the DGA, but the DGA mechanisms for both categories of data are more or less the same.

[27] Council of the European Union, Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act) – Analysis of the final compromise text in view to agreement, 2020/0340(COD), Brussels, 10 December 2021.

> …certain categories of data (commercially confidential data, data subject to statistical confidentiality, data protected by intellectual property rights of third parties, including trade secrets and personal data) in public databases is often not made available, despite this being possible in accordance with the applicable Union law, in particular Regulation (EU) 2016/679 and Directives 2002/58/EC and (EU) 2016/680, not even for research or innovative activities in the public interest. Due to the sensitivity of those data, certain technical and legal procedural requirements must be met before they are made available, not least in order to ensure the respect of rights others have over such data, or limit negative impact on fundamental rights, the principle of non-discrimination and data protection. Such requirements are usually time- and knowledge-intensive to fulfil. This has led to the underutilisation of such data (…) In order to facilitate the use of data for European research and innovation by private and public entities, clear conditions for access to and use of such data are needed across the Union.

The DGA basically introduces three trajectories to incentivise sharing of data that is protected by data protection or intellectual property rights. Firstly, it gives guidance on the re-use of protected data owned by public sector bodies (Chapter II, DGA). Secondly, it creates a "data altruism" framework (Chapter IV, DGA), facilitating the sharing of protected data for the common good, including research. And finally, it introduces a framework for so-called "data intermediation services" (Chapter III, DGA), that is, professional data sharing services. One central idea in the DGA is to create sector-specific "data spaces", which could be described as data silos or commons, managed by the aforementioned data intermediaries. When data are kept in such a data space, supervised and managed by a professional intermediary, this would hopefully lead to improved data quality, reliability, availability and security of data, which would automatically also entail a higher level of GDPR compliance and public trust, as well as a more streamlined and institutionalised process for requesting permission to use the data. Some of these data spaces, such as the European Health Data Space, will require additional regulation[28] because of the specific sensitive nature of certain types of data and particular sectorial demands. The DGA will, however, provide the common

[28] Towards European Health Data Space (TEHDAS), *Milestone 5.8 Potential health data governance mechanisms for European Health Data Space*, 1 September 2021, project report co-funded by the European Union's 3rd Health Programme (2014–2020) under Grant Agreement no 101035467.

framework. In Article 2(c) of the DGA, data intermediation service is defined as 'a service, which aims to establish commercial relationships for the purpose of data sharing between an undetermined number of data subjects and data holders, on the one hand, and data users on the other hand, through technical, legal or other means, including for the exercise of data subjects' rights in relation to personal data'.[29] Data intermediaries will form a new commercial business model which excludes existing non-profit collaborative knowledge platforms (such as *WikiMedia*), as well as commercial businesses that only provide a technical means for sharing without establishing a legal and commercial relationship between potential sharers and users (such as cloud services like *OneDrive* and *Dropbox*):

> The provision of cloud storage, analytics or of data sharing software, the provision of web browsers or browser plug-ins, or an email service should not be considered data intermediation services in the sense of this Regulation, as long as such services only provide technical tools for data subjects or data holders to share data with others, but are neither used for aiming to establish a commercial relationship between data holders and data users, nor allow the provider to acquire information on the establishment of commercial relationships for the purpose of data sharing, through the provision of such services. Examples of data intermediation services would include, inter alia, data marketplaces on which companies could make available data to others, orchestrators of data sharing ecosystems that are open to all interested parties, for instance in the context of common European data spaces, as well as data pools established jointly by several legal or natural persons with the intention to license the use of such pool to all interested parties in a manner that all participants contributing to the pool would receive a reward for their contribution to the pool. This would exclude value-added data services, that obtain data from data holders, aggregate, enrich or transform the data for the purpose of adding substantial value to it and license the use of the resulting data to data users, without establishing a commercial relationship between data holders and data users.[30]

The idea is that data intermediaries would be registered, supervised by a new supervisory body called the 'European Data Innovation Board' and easily recognisable through a common logo that identifies them as a provider of 'data intermediation services recognised in the Union'. As such, these intermediaries, who *only* act as intermediaries and not use the

[29] DGA-Council (n. 28), Article 2c.

[30] DGA-Council (n. 28), Recital 22a.

data themselves for other purposes (Article 2(c) of the DGA), would offer natural and legal persons an alternative to simply parking their data at some integrated tech platform. Storing data at an intermediation service would offer a way for data subjects and data holders to stay in control of the data connected to them, through data protection or intellectual property rights, while at the same time allowing for data sharing for certain purposes.

One important adjustment in the latest version of the DGA is a clarification in Recital 3a, namely that the GDPR has an incontestable primacy over the DGA:

> This Regulation should in particular not be read as creating a new legal basis for the processing of personal data for any of the regulated activities, or as modifying information requirements under Regulation (EU) 2016/679.[31]

Thus, if the intermediation services are to make data sharing easier, this would not be because the regulatory data protection regime is altered. The introduction of intermediation services in the DGA aims to be like a *Tinder* of sharing data. Without altering the data protection rules as such, the hope is that the introduction of data intermediaries, who match potential data sharers with potential data (re-)users, could be as much of a game changer as *Tinder*-like services were for dating.

In order to further incentivise natural and legal persons to share data with data, the EU legislator has also introduced the concept of 'data altruism' in Article 19 DGA. In contrast to data intermediaries, data altruism organisations and the natural or legal persons sharing their data for altruistic purposes do this on a *non-profit* basis. Data altruism organisations should, like data intermediaries, be recognisable by a common logo. Article 2(10) of the DGA states that 'data altruism' amounts to 'voluntary sharing of data based on consent by data subjects to process personal data pertaining to them, or permissions of other data holders to allow the use of their non-personal data without seeking or receiving a reward that goes beyond a compensation related to the costs they incur making their data available, for purposes of general interest'. As pointed out by González Fuster,[32] the choice of the word 'data altruism', instead of, for example, the more neutral term 'data donation', gives a strong normative value to

[31] DGA-Council (n. 28), Recital 3a.

[32] Gloria González Fuster, Carta Academica. L'altruisme des données peut-il sauver le monde? Le Soir, 24 April 2021.

the concept. ‘Data altruism’ has a morally positive ring to it, whereas its opposite, ‘data egoism’, sounds less appealing. Moreover, González Fuster continues, the problem with the word ‘data donation’ is also that data protection is understood in the EU as an inalienable fundamental right that should not be understood in terms of property rights. One cannot sell or give away one’s right to data protection, in the same way as one cannot do that with other inalienable rights, such as one’s right to human dignity. The word ‘data altruism’ makes it easier to defend that data are not donated, in the meaning of a transfer of property, but that the data subject consents to its use in compliance with the GDPR. Not unlike the data spaces managed by commercial intermediaries, data altruism organisations fulfil the role of a dating market between potential data sharers and users, but what brings them together is a non-commercial shared commitment to a particular purpose of general interest. Article 22 of the proposed DGA offers the possibility to the Commission to adopt implementing acts for the development of a uniform European data altruism consent form, using a modular approach, allowing customisation for specific sectors and for different purposes. This consent can, in line with the GDPR, be revoked at any point. It could, however, be questioned if the data altruism in the proposed DGA is as fully GDPR compatible, as it claims to be.[33] In its position paper[34] on the DGA, the European Consumer Organisation (BEUC) warns that the term ‘purposes of general interest’ (Article 2(10) of the DGA) is too vague. The term can easily be stretched in unforeseeable ways:

> Consumers must also be legally protected against misleading practices which are presented as public purpose research when in reality there is commercial intent in the exploitation of the data as a result of the commercialisation of the research outputs.[35]

[33] Paul Keller and Francesco Vogelezang, The Data Governance Act – between undermining the GDPR and building a Data Commons, *EDRI*, at: https://edri.org/our-work/the-data-governance-act-between-undermining-the-gdpr-and-building-a-data-commons/ (published online 14 July 2021); Paul Keller and Francesco Vogelezang, The Data Governance Act: five opportunities for the data commons, *Open Future*, at: https://openfuture.eu/publication/the-data-governance-act-five-opportunities-for-the-data-commons/ (published online 23 June 2021).

[34] The European Consumer Organisation (BEUC), *Data Governance Act. BEUC position paper*, 2021.

[35] BEUC (n. 35), p. 3.

In the initial version of the DGA, 'general interest' was left undefined and only exemplified by two examples in Article 2(10): '…such as scientific research purposes or improving public services'. BEUC criticised this lack of a definition of 'general interest' in the DGA and wrote that:

> there are no clear legal benchmarks to check against the presence of such a 'general interest' ('altruism washing') and, in some cases, the interpretation of what constitutes a 'general interest' might differ at national level.[36]

In the latest version of the DGA, 'general interest' is illustrated with more examples in Article 2(10); nonetheless, at a fundamental level, BEUC's criticisms regarding vagueness still hold:

> …for purposes of general interest, defined in accordance with national law where applicable, such as healthcare, combating climate change, improving mobility, facilitating the establishment of official statistics, improving public services, public policy making or scientific research purposes in the general interest.

It would be up to altruism organisations or the data receiving public entity to ensure that the altruistically shared data are shared for a purpose or set of purposes that can be qualified as 'general interest' and that are sufficiently specific to be in accordance with the purpose specification principle in Article 5(1)(b) of the GDPR. The importance of compliance with the purpose specification principle is clarified by a new addition in Article 19(1)a. In the initial version of this Article, it only said that data altruism organisations should inform data holders 'about the purposes of general interest for which it permits the processing of their data by a data user', whereas the latest version also specifies that information about 'the specified, explicit and legitimate purpose' for which it permits the processing of personal data should be provided. However, the fact that a data space, managed by a data intermediation service or data altruism organisation, is supposed to be a one-stop shop, where a multitude of actors can request access to data, seems to create an incentive to not make the purposes too specific, and makes it attractive to stretch out the specificity of purposes to the maximal vagueness still permitted by the GDPR. Moreover, in the case of data altruism, one could basically imagine two different scenarios with regard to purpose specification. The first one is

[36] BEUC (n. 35), p. 8.

where potential data sharers have a very specific purpose in mind, such as in the case where a patient suffering from a rare disease wants to stimulate research only in this very particular field. The second scenario is a potential data sharer who simply wants to get a quick fix of 'do-gooder' feeling and is nudged to share data for a default set of rather broadly formulated general interest purposes. The question is if the latter would cause friction with the requirements of purpose specification and freely given consent in the GDPR. Moreover, data altruism might give the false impression to data subjects that re-use of data for a new purpose always requires their renewed consent (Article 6(4) of the GDPR), whereas, in fact, this is not the case for 'archiving purposes in the public interest, scientific or historical research purposes or statistical purposes' (Article 5(1)b)) that are presumed to be compatible, as well as for re-use that is in accordance with the law, necessary in a democratic society and in pursuance of a legitimate aim (Article 6(4) of the GDPR).[37] A potential do-gooder might be surprised to find out that a planned act of altruism is void because the data already have been shared, and that the GDPR, in fact, does not always require consent for data re-use.

Does the DGA help a researcher who is experiencing GDPR anxiety? The DGA might help, in terms of data *accessibility*, in the same way a dating service like *Tinder* increases the amount of potential individuals to date. However, given that the GDPR has primacy over the DGA, the burden of GDPR compliance will not disappear. The procedures followed by data intermediation services and data altruism organisations might be more standardised, but the substantive thinking about data protection requirements cannot be removed. Nor are GDPR requirements like data minimisation and purpose specification, which might frustrate a researcher who would have the freedom to freely change between research purposes: if a data set containing gait and facial expressions of individuals in public transport does not lead to a good AI-model to identify Covid-19 infections, why not try to see if the data can be used to spot people who don't have a bus ticket or illegal migrants? Even though the GDPR, in principle, allows for jumping from one research purpose to another (compatible purpose, Article 5(1)b of the GDPR), the researcher would have to do the exercise in substantive GDPR-thinking before each shift in research purpose. In order to truly stop worrying about the

[37] Merel Koning, *The purpose and limitations of purpose limitation*, Doctoral dissertation Radboud University, 2020.

GDPR, a researcher would have to find data that fall outside the scope of the GDPR. One possible route to do this is by using anonymous data. Sometimes, training an AI-model on anonymised data is a viable option, but sometimes anonymisation of personal data leads to too much utility loss. The holy grail is then to find a surrogate to personal data that has the same utility yet does not qualify as personal data. One possibility – at least in most Member States – is to use data of deceased people (see section 4 below). Another one is to use so-called synthetic data, which are fake data that resemble real personal data. (see section 5 below).

4 A surrogate to personal data: Data of deceased people

Recital 27 explicitly states that the GDPR does not apply to the personal data of deceased persons, but that Member States may provide for rules regarding the processing of personal data of deceased persons. Approximately two-thirds of EU Member States have not chosen to do so.[38] For example, The Netherlands and Sweden have not created any provisions for data of deceased individuals; however, in Denmark, there is 10 years of protection after moment of death (§ 2(5) Danish Data Protection Act[39]). This means that in many Member States, data of deceased individuals could be a legal loophole and a window of opportunity for certain types of research. There are, however, several caveats to take into account.

Firstly, data of deceased people only fall outside the scope of the GDPR if they do not relate to any living individual.[40] A post on a social network containing information about both a deceased and a living individual would still qualify as personal data in the meaning of the GDPR. Certain types of data, such as genetic data, almost always also relate to living people even if they are primarily related to a deceased individual.

[38] David Erdos, Dead ringers? Legal persons and the deceased in European data protection law, 40 Computer Law & Security Review 40, 2021. See for an overview, for example: https://www.twobirds.com/en/in-focus/general-data-protection-regulation/gdpr-tracker/deceased-persons (last accessed 10 December 2021).

[39] Act No. 502 of 23 May 2018, published in the Law Gazette on 24 May 2018, at: https://www.datatilsynet.dk/media/7753/danish-data-protection-act.pdf.

[40] Iñigo de Miguel Beriain, Aliuska Duardo-Sánchez, José Castillo Parrilla, What Can We Do with the Data of Deceased People? A Normative Proposal, 29 European Review of Private Law, Issue 5 (2021), pp. 785–806.

Secondly, certain types of research need to be approved by a research ethics body. It should be noted that data protection and research ethics depart from different legal rationales: the former is strongly connected to informational self-determination, while the latter connects more to human dignity. National guidelines and regulations on research ethics often differ quite substantially from one country to another. However, some widely recognised international codes exist. One of the most important ones is the World Medical Association's Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.[41] The first paragraph of its Preamble, the Helsinki Declaration, says that it applies to 'medical research involving human subjects, including research on identifiable human material and data'. Medical research involving human participants, genetic data or human tissue are classical types of research that are required to undergo ethical review in almost any country. Research from other domains, such as humanities[42] or natural sciences,[43] can sometimes also be required to undergo ethical review. In Sweden, the *Act concerning the Ethical Review of Research Involving Humans*[44] makes a link in 3§(1) to sensitive personal data, as defined in Article 9 of the GDPR: any research processing such data has to apply for ethical approval. It should, however, be underlined that despite the link to the GDPR research, ethics assessments follow their own logic that has to be clearly distinguished from an assessment of data protection compliance. Research ethical assessments often, for example in the aforementioned Swedish Act and the Helsinki declaration, have two main elements: independent ethical oversight that balances the scientific value against the privacy, health and safety it may entail for involved human participants' risks (and where priority is given to the latter) and informed consent. From a research ethics perspective, obtaining informed consent from study participants, unless they are deceased, is almost always necessary. This should be contrasted with the data protection law, where consent is only one

[41] World Medical Association's Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, adopted by the 18th WMA General Assembly, Helsinki, Finland, June 1964. Latest amendment on the 64th WMA General Assembly, Fortaleza, Brazil, October 2013.

[42] Ulf Görman, *Lathund för etikprövning – Humanistiska och teologiska (HT) fakulteterna*, Lunds Universitet, 2017.

[43] Etikprövning – en översyn av reglerna om forskning och hälso- och sjukvård (SOU 2017:104).

[44] Lag om etikprövning av forskning som avser människor, SFS 2003:460.

of several legal grounds for processing (Article 6(1) of the GDPR) and an enormous amount of processing happens on other grounds, without consent, such as the public or legitimate interest grounds. Thus, the 'freely-given, specific, informed and unambiguous' consent in data protection law differs 'conceptually and operationally'[45] from the informed consent that research ethics requires, the former rooted in information self-determination and the latter in human dignity. This distinction also explains why research ethics codes often do not exclude data relating to deceased people in the same way as the GDPR. For example, the use of biological material of a deceased human being might encroach on post-mortem human dignity, but informational self-determination is no longer applicable if an individual is not alive.

Thirdly, the question is if the right to private life in Article 8 of the European Convention on Human Rights (ECHR) could cause problems for the use of data relating to deceased people. This is rather unlikely. Despite the fact that the Strasbourg Court does not fully exclude that Article 8 can be applicable to deceased individuals,[46] the main focus is clearly on living people.

Finally, one should always take into consideration if data relating to deceased people are protected by some other rights of others, such as copyright, database rights or other intellectual property rights. This could, for example, be relevant with regard to pictures of deceased people that qualify as copyright protected works.

In summary, deceased people's data could be a good alternative for researchers who want to escape the scope of the GDPR, as long as the data does not relate to other living beings, and potentially applicable national data protection legislation about deceased people, research ethics and intellectual property laws are taken into consideration. Krutzinni and Floridi[47] have proposed that the use of medical data of deceased people should be facilitated by creating a dedicated medical code for posthumous data donation (PMDD) that enables individuals to decide how their medical data could be used after their death, in a manner akin to

[45] EDPS, Opinion on scientific research (no 2), 2.

[46] European Court of Human Rights (ECtHR), *M. L. v Slovakia* (Application no. 34159/17), 14 October 2021.

[47] Jenny Krutzinna & Luciano Floridi, Ethical Medical Data Donation: A Pressing Issue, in Jenny Krutzinna & Luciano Floridi, *The ethics of medical data donation*, Springer Nature, 2019, 1–6.

how one decides 'to donate blood, organs or tissue'.[48] While this proposal for a dedicated PMDD code has not been adopted by any legislator yet, it seems to fit in well with the spirit of the aforementioned DGA, and it is not difficult to see how data altruism could be stretched out to apply to this kind of posthumous data altruism too. While Krutzinna, Taddeo and Floridi[49] consider it ethically preferable to begin with creating a framework for posthumous data donation and later possibly extend to medical data donation by living individuals and corporations, the EU legislator seems to work from the other direction by introducing data intermediation services and data altruism in the proposed DGA.

5 A surrogate to personal data: Synthetic data (a particular type of anonymised data)

The material scope of the GDPR, as discussed above (in section 2), is very broad because of the enormous amount of data that fall under the definition of personal data, as defined in Article 4(1) of the GDPR.[50] This means that many researchers face GDPR questions, unless they find a way to escape its scope. Using data of deceased people as a way to escape the scope of the GDPR is only a minor fringe phenomenon compared to the most classical way to do so, namely by using non-personal data or by anonymising data.

Apart from the fact that national complementary provisions can make data protection extend to data of deceased people, such data also have practical limitations. Not all research can be based on data relating to deceased people. For example, in order to create AI models that capture contemporary phenomena, such as symptoms caused by the latest variety of the Covid-virus, outdated data of deceased people will not do. If data of deceased people will not help a researcher, it might be time to look at the more conventional road: anonymisation.

[48] Krutzinna & Floridi (no 47), 2.

[49] Jenny Krutzinna, Mariarosaria Taddeo, and Luciano Floridi, Enabling Posthumous Medical Data Donation: A Plea for the Ethical Utilisation of Personal Health Data, in Jenny Krutzinna & Luciano Floridi, *The ethics of medical data donation*, Springer Nature, 2019, 163–180.

[50] GDPR 2016/679 (n. 3).

In order to anonymise personal data, their link to any 'identified or identifiable natural person' (Article 4(1) of the GDPR[51]) needs to be severed in a way that cannot be reversed with 'all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly' (Recital 26 GDPR). This is not an easy feat. Removing a name or personal identification number is hardly ever enough to anonymise the data properly because the remaining data in a profile might be rich enough to single out a particular individual. For example, a data profile that refers to a female academic with Dutch origin working in public law at Uppsala University singles me out, despite the fact that it does not contain my name or some other unique identifier. Re-identification is often facilitated by the combination of different data sets or by using novel data techniques. This means that seemingly anonymised data, in practice, often actually should be qualified as pseudonymised data, defined in Article 4(5) of the GDPR as data that can 'no longer be attributed to a specific data subject without the use of additional information'. Pseudonymised data are a particular type of personal data, and thus still fall in the scope of the GDPR. This means that often, in order to realise true anonymisation[52] in the GDPR, quite a substantial amount of information should be removed, which can lead to a loss of utility. This is what is commonly known as the privacy-utility trade-off: by removing information, data might become anonymised, but this is of little avail if it disfigures the data to such an extent that they are no longer useful for the intended research. The holy grail of anonymisation is thus to find techniques that prevent re-identification while preserving data utility. During the last few years, synthetic data[53] have been proposed as potentially being this holy grail of anonymisation.[54] The basic idea is that instead of removing data, an AI model is trained on real data to generate fake data with the same statistical properties. An example would be to create a generative model that creates convincingly realistic portrait pictures of non-existing peo-

[51] GDPR 2016/679 (n. 3).

[52] Even though written when the GDPR was not in force yet, many of the arguments are still applicable: Working Party 29, *Opinion 05/2014 on Anonymisation Techniques*, 2014.

[53] Luke Rodriguez & Bill Howe, In Defense of Synthetic Data, arXiv:1905.01351, 2020; Anjana Ahuja, The promise of synthetic data, Financial Times, 2020; Laboratory for Information and Decision Systems, The real promise of synthetic data, MIT News, 2020.

[54] Steven M. Bellovin, Preetam K. Dutta, Nathan Reitinger, Privacy and Synthetic Datasets, 22 Stanford Technology Law Review, Issue 1, 2019, 1–51.

ple[55] and to use these simulated data as a basis to train an AI model. The question is, of course, if training a model on synthetic data will generate sufficiently good results in comparison to using real data. Some authors[56] have claimed that this might not be the case and that synthetic data of good quality might still be traceable to identifiable or identified individuals, and that the privacy-utility trade-off is not truly resolved by using synthetic data. Despite the drawbacks, using synthetic data might be a viable option for at least some types of relatively simple data (for example, a portrait photo might be more easily simulated than a brain scan) with little outliers (for example, if a data set of portrait photos contains 100,000 human faces and 5 cat faces, the cat pictures in the synthetic data set will probably be much closer to the original pictures than the human ones, simply because there has not been enough material to generalise).

6 Conclusions: There is no one size that fits all

What to tell a researcher who needs data to build an AI-model but fears that the GDPR-requirements will create a burden that is too large and too demanding? The first message is a comforting one. The GDPR has a broad understanding of scientific research, and it has a rather generous research exception in Article 89 of the GDPR. Nevertheless, GDPR compliance is more than just ticking a few boxes and will often require some substantive thinking about data protection risks and balancing of different interests. The EU legislator partially helps researchers in the proposed DGA, by creating infrastructures that will help match potential data sharers and data users. Data intermediation services and data altruism organisations are thus likely to increase data access to data that are protected by rights of others (data protection or intellectual property rights). However, it should be underlined that the GDPR has primacy over the DGA and that the improvements are mostly in terms of data availability and infrastructure. The potential burden of compliance with GDPR requirements is not altered by the proposed DGA. For researchers that want to use data and not be burdened by the GDPR, I discuss two alternatives. Firstly, one could consider using data of deceased individuals, in as far as they are not connected to any other living individuals.

[55] See https://thispersondoesnotexist.com/ (last accessed 10 December 2021).

[56] Theresa Stadler, Bristena Oprisanu & Carmela Troncoso, Synthetic Data – Anonymisation Groundhog Day, arXiv:2011.07018, 2022.

Here, also other legislative instruments should be taken into account that could potentially limit the ways in which such data may be used: national data protection provisions on data of deceased individuals, research ethics codes and intellectual property laws. Secondly, one could consider using anonymised data. In those cases where traditional anonymisation methods degrade the utility of the data too much, the use of synthetic data could be an option.

In summary, the researcher suffering from GDPR-anxiety, who is looking for personal data or surrogates with a similar level of utility, in principle, has a *smörgåsbord* of options to pick from. However, which type of data will be helpful in a particular research project is a highly contextual question – in finding the right type of research data, there is no *one size fits all*.

Bengt Domeij

Krav på att få nyttja andras industriella data (som inte är personuppgifter) för att kunna utveckla AI-tjänster: en översikt

1 Inledning

Samhället utvecklas och förändras genom att vissa tillgångar blir mer värdefulla medan andra tappar i betydelse. De senaste åren har ofta sagts att data är vår tids olja.[1] Med data har då främst åsyftats personuppgifter. Annan data än personuppgifter har inte uppmärksammats i samma utsträckning, förmodligen eftersom den ännu inte har fått lika stor praktisk användning, exempelvis för riktad reklam, och inte heller ger särskilda integritetsproblem. Industriell data, eller icke-personuppgifter, är något som huvudsakligen insamlas i realtid genom sensorer på produkter och väntas få allt större betydelse de närmaste åren. Industriell data är bland annat värdefullt vid utveckling av AI-baserade tjänster för optimering av olika slags processer, allt ifrån trafikflöden till jordbruk och industritillverkning.[2] För att träna en AI-algoritm som kan styra en sådan process,

[1] Se exempelvis https://www.economist.com/ May 6th 2017 edition, artikeln ”The world's most valuable resource is no longer oil, but data”. Se vidare A. Nordberg, Trade Secrets, Big data and Artificial Intelligence Innovation: a Legal Oxymoron?, The Harmonization and Protection of Trade Secrets in the EU: An Appraisal of the EU Directive, (red.) J. Schovsbo, T. Minssen, T. Riis (2020) s. 194.

[2] Preambel 9, Förordning (EU) 2018/1807 om en ram för det fria flödet av andra data än personuppgifter i Europeiska unionen: ”Det växande sakernas internet (IoT), artificiell intelligens och maskininlärning utgör viktiga källor till andra data än personuppgifter, till exempel som ett resultat av deras användning inom automatiserade industriella pro-

behövs som regel mycket stora mängder data.[3] Stundom är det samma aktör som har tillgång till data och som planerar att utveckla en AI-baserad tjänst, men det är långt ifrån alltid fallet.[4] Det är också så att aggregerad data ofta är värd mer än summan av de ingående datasamlingarna.[5] Det sagda riktar uppmärksamheten mot ett problem: datasamlingar för träning av AI förflyttas inte enkelt dit där de ger mest värde.

Data är ett objekt med särskilda avtalsproblem för företag och innovatörer. Det är ett förhållandevis nytt avtalsobjekt och kan ha flera olika och ibland svårinsedda användningsmöjligheter. En biltillverkare kan via sensorer ha samlat in data för att underlätta service av bilen, men samma data kan kanske nyttjas för utveckling av system för självkörande bilar, men också av försäkringsbolag för att anpassa förarens försäkringspremium. Det finns förmodligen andra och mer svårinsedda användningsmöjligheter för information i realtid om hur enskilda bilar framförs. Ta idén om en tjänst som anger tidsåtgång för alternativa färdvägar och som bygger på rådande genomsnittshastighet på olika vägar. En utvecklare av en sådan tjänst skulle förmodligen behöva data från många av de viktigaste fordonstillverkarna för att kunna utveckla tjänsten. Detta skulle vara svårt att avtala om och kanske hindra den aktuella tjänsten. Ett särskilt problem vid nya AI-tjänster är att en innovatör eventuellt inte vågar avslöja sin avsikt beroende på risken för att datainnehavaren själv väljer att utnyttja möjligheten. En patentansökan brukar användas som ett sätt att skaffa sig en säkrare förhandlingsposition när ett utvecklingsprojekt ska inledas i samarbete med andra.[6] Men det är svårt att få immaterialrättsligt skydd för en idé om en ny tjänst. Innan idé prövats experimentellt är den

duktionsprocesser. Konkreta exempel på andra data än personuppgifter inkluderar aggregerade och anonymiserade datamängder som används för stora dataanalyser, data om precisionsjordbruk som kan bidra till övervakning och optimering av användningen av bekämpningsmedel och vatten, eller data om underhållsbehov för industriella maskiner.”

[3] E. Rosati, Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity, elektroniskt tillgänglig på https://ssrn.com/abstract=3452376.

[4] A. Wiebe, Protection for Industrial Data – A New Property Right for the Digital Economy? Journal of Intellectual Property Law & Practice, 12(1), p. 63 (2017).

[5] L. Cabral, et al, The EU Digital Market – A Report from a Panel of Economic Experts s. 20, The European Commission science and knowledge service, 2021, tillgänglig på https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3783436.

[6] B. Domeij, Patent och innovationsprocessens avtal s. 59–75 i Ett innovationspolitiskt ramverk - ett steg vidare, (red. P. Braunerhjelm), Entreprenörskapsforum, 2011, och även publicerad i NIR 2012 s. 122–140.

knappast patenterbar, eftersom tjänstidén i sig inte löser något tekniskt problem. Det finns knappast heller någon annan immaterialrätt som står till förfogande. Sammanfattningsvis är industriell data ett nytt värdefullt objekt som har särskilda svårigheter vid samarbetsavtal. När data inte kan överföras blir resultatet datasilos och att den fulla potentialen hos nya AI-tjänster riskerar att gå förlorad.

Den europeiska utvecklingen kring industriell data inledes med en diskussion av om det skulle behövas ett nytt immaterialrättsligt skydd.[7] Införandet av en äganderätt underlättar, som sagt, avtal och är ett vanligt övervägande när ett immateriellt objekt genom teknisk utveckling har fått ökad betydelse. EU-kommissionen prövade i början av 2017 idén om att skapa en immaterialrätt för industriell data.[8] I bland annat akademiska inlägg argumenterades emellertid för att det saknades ett sådant behov och att risken snarast var att den faktiska kontroll som fanns hos företag ifråga om industriell data, som uppnåddes genom tekniska spärrar mot andras åtkomst och skyddet som finns för företagshemligheter, redan innebar att delningen mellan företag av icke-personuppgifter var alltför begränsad.[9] Det bedömdes vara osannolikt att det fanns ett reellt behov av ytterligare incitament för utveckling av industriell data eller för nya begränsningar ifråga om användning. Industriell data uppstår, som sagt, genom sensorer som byggs in i produkter i syfte att kunna styra och övervaka funktionen. Kostnaderna för detta är inte jämförbara med de som finns vid ett större forskningsprojekt där utfallet brukar vara högst osäkert. Och efter kort tid kom diskussionen att istället gälla regler för ökad delning av data, både frivillig och tvingande sådan.[10] EU-kommissionens nya bild av läget för icke-personuppgifter år 2020 var att det fanns

[7] M. Leistner, The existing European IP rights system and the data economy – An overview with particular focus on data access and portability, s. 249 i Data Access, Consumer Interests and Public Welfare, tillgänglig på https://doi.org/10.5771/9783748924999-1, am 14.07.2021 och https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3625712.

[8] Communication from the European Commission of 10 January 2017 – Building a European Data Economy, COM(2017) 2 final, s. 13.

[9] J. Drexl, Designing Competitive Markets for Industrial Data, Max Planck Institute for Innovation and Competition Research Paper No. 16-13, tillgängligt på https://ssrn.com/abstract=2862975.

[10] J. Drexl, Connected devices – An unfair competition law approach to data access rights of users, DOI:10.5771/9783748924999-477, publicerad i boken Data Access, Consumer Interests and Public Welfare, s. 477 (483), och tillgänglig på https://www.researchgate.net/publication/350522751_Connected_devices_-_An_unfair_competition_law_approach_to_data_access_rights_of_users.

en bristande tillgång hos företag och avtalsproblem som kunde hindra innovationer i form av nya tjänster baserade på AI.

> Värdet av data ligger i deras användning och vidareutnyttjande. För närvarande finns det inte tillräckligt med data tillgängliga för innovativt vidareutnyttjande, såsom utveckling av artificiell intelligens. Problemen kan grupperas enligt vem som innehar respektive använder uppgifterna, men också utifrån vilken typ av data det rör sig om (dvs. personuppgifter, icke-personuppgifter eller blandade datamängder med både och).[11]
>
> Trots den ekonomiska potentialen har datadelning mellan företag inte tagit fart i tillräcklig utsträckning. Detta beror på avsaknaden av ekonomiska incitament (och rädslan för att förlora en konkurrensfördel), ekonomiska aktörers bristande tilltro till att uppgifterna kommer att användas avtalsenligt, ojämna förhandlingspositioner, rädslan för att tredjeparter ska tillskansa sig uppgifterna samt en brist på rättslig klarhet om vem som kan göra vad med data (t.ex. med gemensamt skapade data, framför allt data från sakernas internet).[12]

EU Kommissionens avsikt är nu att förbättra företags tillgång till industriell data genom ett förslag till en ny Data Act som ska presenteras under fjärde kvartalet 2021.

> [Ett förslag till en ny EU Data Act] will aim to increase access to and further use of data, so that more public and private actors can benefit from techniques such as Big Data and machine learning. The conditions of access and further usage in B2B relationships are often regulated by private contracts. The initiative would look both at data usage rights in industrial value chains and particularly at a fair distribution of usage rights that allow all parties to benefit from data-driven innovation. … More specifically in this respect, the assessment will consider the following problems:
>
> i. B2B data sharing works best where the data holder has an incentive to share data and the parties' negotiating power is comparable. A data holder with a stronger negotiating power may, however, unilaterally impose unfair terms and conditions to the detriment of a company seeking data access which could have the effect of making data sharing disproportionately difficult or economically prohibitive or refuse access to data altogether. This may prevent data-driven businesses from developing/running their business models and could push existing market players out of the market and prevent new players from entering the market.

[11] En EU-strategi för data s. 6, från den 19.2.2020, COM(2020) 66 final.

[12] En EU-strategi för data s. 7, från den 19.2.2020, COM(2020) 66 final.

> ii. Non-personal data co-generated through industrial use constitute a specific class of data that will grow at exponential scale over the years to come (factory robots, agricultural machinery, etc.). The attribution of the rights to access and use such data is left to private contract. This can raise questions of fair competition in terms of the different markets (supplier, OEM-buyer relations, aftermarkets). Also, there is untapped innovative potential in secondary and tertiary uses made of the data through the development of novel services that rely on access to such data.[13]

Grundproblemet är avtalssvårigheter som kan hindra innovation, något som är särskilt påtagligt för mindre och medelstora företag. Vid sådana förhållanden är naturligtvis framtvingad delning ett alternativ. En översikt av när så sker med icke-personuppgifter i europeisk rätt är ämnet för denna artikel. När det nedan talas om tvingande datadelning avses och inkluderas däri möjligheten för mottagaren att utföra de handlingar med data som mottagaren önskar, exempelvis att träna en AI-algoritm; ämnet för artikeln är således ”access and use” ifråga om icke-personuppgifter. Rättsmedlen i Europa som har detta syfte är dels sektorspecifik lagstiftning, dels konkurrensrätt. Först ska emellertid kort utvecklas vad det praktiskt innebär att industriell data skyddas som företagshemligheter.

2 Industriell data som företagshemligheter

Regler kring företagshemligheter har bred tillämpning och omfattar all slags teknisk och kommersiell information hos näringsidkare, t.ex. kundregister och testdata. Åtminstone den övervägande mängden data hos företag som kan bli aktuell vid utveckling av AI-algoritmer uppfyller kriterierna i 2 § LFH (lag 2018:558 om företagshemligheter).[14] Data inhämtad från sensorer placerade på industriella produkter är inte, varken som helhet eller i den form dess beståndsdelar ordnats och satts samman, allmänt känd hos eller lättillgänglig för de som normalt har tillgång till

[13] EU Kommissionens Inception Impact Assessment, tillgängigt på https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13045-Data-Act-including-the-review-of-the-Directive-96-9-EC-on-the-legal-protection-of-databases-_en.

[14] Den svenska definitionen av företagshemligheter härrör från direktiv 2016/943 om skydd mot att icke röjd know-how och företagsinformation (företagshemligheter). Se närmare A. Nordberg, Trade Secrets, Big data and Artificial Intelligence Innovation: a Legal Oxymoron?, The Harmonization and Protection of Trade Secrets in the EU: An Appraisal of the EU Directive, (red.) J. Schovsbo, T. Minssen, T. Riis (2020) s. 194.

information av det aktuella slaget. Informationen är unik för företaget. Vidare skulle ett röjande, så som krävs i 2 § LFH, vara ägnat att medföra skada i konkurrenshänseende för innehavaren. Slutligen, ska innehavaren ha vidtagit rimliga åtgärder för att hemlighålla databasen, t.ex. att anställda eller affärspartners genom exempelvis lösenord eller avtal förstår eller förutsätter att innehavaren har avsett att hemlighålla informationen.[15] Att informationen hade kunnat samlas in av andra, saknar betydelse för frågan om det finns ett skydd som företagshemligheter, så länge som data inte offentliggjorts av den som samlat in information. Delad data har fortsatt skydd, givet att mottagarna förstår att det rör sig om hemligheter och att de inte får spridas vidare. Företagshemligheter ska kunna spridas kontrollerat.[16]

Lagen skyddar mot att någon olovligen bereder sig tillgång till företagshemligheter, exempelvis tar sig förbi tekniska skydd. För den som lovligt har mottagit annans industriella data, sker ett olovligt utnyttjande om mottagaren använder företagshemligheterna i strid mot avtalet. Om exempelvis en mottagare av data har fått tillstånd att utveckla en viss AI-baserad tjänst, men istället utvecklar en annan AI-baserad användning, är handlandet i strid mot 3 § första stycket LFH. Om en mottagare har licens att använda data för eget nyttjande, men låter någon annan ta del av informationen, innebär det ett otillåtet röjande, enligt 3 § första stycket LFH.

I lagen om företagshemligheter från 2018 infördes regeln om att produkter som utomlands tillverkats på ett sätt som gynnats avsevärt av ett olovligt utnyttjande av företagshemligheter utanför Sverige, innebär intrång i de svenska företagshemligheterna när slutprodukterna förs till Sverige (utan att slutprodukten i sig innehåller företagshemligheterna, 3 § andra stycket LFH). Det är tveksamt om regeln träffar en AI-tjänst som utomlands tagits fram genom ett otillåtet utnyttjande av data/företagshemligheter. Enligt ordalydelsen i lagtexten gäller förbudet importerade varor, som exempelvis tagits fram genom någon annans hemlighållna tillverkningsprocess. Det finns, så vitt jag kan se, inget egentligt skäl till att en AI-baserad tjänst som utomlands tagits fram genom intrång i, i Sverige skyddade företagshemligheter, bör få erbjudas på den svenska marknaden. Men lagstiftning av detta slag bör inte tolkas i strid mot ordalydelsen. Det förefaller därmed finnas en begränsning i skyddet

[15] Prop. 2017/18:200 En ny lag om företagshemligheter s. 31.

[16] B. Domeij, Från anställd till konkurrent s. 128 (2016).

för företagshemligheter genom att hemlighållen data olovligen kan användas i utlandet för att utveckla och sedan erbjuda en ny tjänst i Sverige utan att den rättmätige innehavaren av företagshemligheterna kan hindra erbjudandet.

I tillägg till skyddet som företagshemligheter, finns det så kallade databasskyddet. Det är emellertid inte så relevant för industriell data, som man kanske skulle tro.[17] Databasskydd innebär dels en möjlighet till upphovsrättsligt skydd för en originell struktur på en databas, dels finns ett *sui generis*-skydd för kopiering av sådana delar av en databas som förutsatt en väsentlig investering vid insamlingen av data.[18] Vid utveckling av AI-tjänster finns behov av data, men värdet ligger knappast i den struktur som data har. Det gör ett eventuellt upphovsrättsligt skydd för strukturen relativt ointressant. De för AI-utveckling aktuella databaserna är normalt inte allmänt tillgängliga, vilket betyder att inte heller *sui generis*-skyddet för databaser spelar någon större roll. *Sui generis*-skyddet är främst väsentligt vid offentliggjorda databaser.[19] Vidare kan påpekas att det förmodligen ofta inte har krävts en väsentlig investering för data som samlats genom ett företags produktsensorer.[20] Man kan konstatera att ett eventuellt databasskydd åtminstone inte framstår som centralt. Nu aktuell data hos företag är tekniskt åtkomstblockerad och utgör företagshemligheter, vilket räcker.[21]

Sammanfattningsvis finns ett skydd, vilket hindrar angrepp på sådan data som kan användas av företag för utveckling/träning av AI-baserade tjänster. När företag, genom sensorer, samlar in data uppstår företagshemligheter. Skyddet gäller utan behov av särskilda åtgärder och omfattar olovligt anskaffande, utnyttjande och röjande av data.

[17] J. Drexl, Designing Competitive Markets for Industrial Data s. 20, Max Planck Institute for Innovation and Competition Research Paper No. 16-13, tillgängligt på https://ssrn.com/abstract=2862975.

[18] Direktiv 96/9/EC från 11 mars 1996 om rättsligt skydd för databaser. Se J. Axhamn, Databasskydd s. 121 ff (2017).

[19] Se J. Axhamn, Databasskydd s. 221 ff (2017).

[20] M. Leistner, The existing European IP rights system and the data economy – An overview with particular focus on data access and portability, s. 223 i Data Access, Consumer Interests and Public Welfare, tillgänglig på https://doi.org/10.5771/9783748924999-1, am 14.07.2021 och https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3625712.

[21] B. Domeij, Databasskydd och företagshemligheter, i G. Karnell, A. Kur, P-J. Nordell, D. Westman, J. Axhamn, S. Carlsson (ed.), Liber Amicorum Jan Rosén s. 254 (2016).

3 Lagstadgad tvingande datadelning

Det är drastiskt att tvinga företag att dela data/företagshemligheter. Det är därför knappast förvånande att EU-kommissionen har uttalat att en skyldighet för företag att dela insamlad data inte kan vara grundregeln i europeisk rätt, men samtidigt har sagt att en sådan skyldighet kan införas om tre villkor är uppfyllda: (1) skyldigheten till datadelning är begränsad till specifika sektorer/branscher, (2) det har konstaterats att det inte är möjligt att få till stånd frivillig handel med data i det aktuella fallet, (3) konkurrensrätten inte kan tillämpas för att lösa problemet.[22] Kriterierna anses vara uppfyllda – tvång att dela data existerar – i ett fåtal specifika fall. En första situation är fordonsbranschen och vid reparations- och servicetjänster.

> Förordning 715/2007 om typgodkännande av motorfordon … och om tillgång till information om reparation och underhåll av fordon, som stadgar i artikel 6.1 att: ”Tillverkarna skall utan dröjsmål ge oberoende aktörer obegränsad och standardiserad tillgång till information om reparation och underhåll av fordon via lättillgängliga webbplatser och i standardiserat format och på ett sådant sätt att de inte diskrimineras jämfört med auktoriserade återförsäljare och verkstäder i fråga om den tillgång och information de sistnämnda ges. För att detta mål lättare skall kunna nås skall informationen lämnas på ett konsekvent sätt, inledningsvis i överensstämmelse med Oasis-formatets tekniska krav. …”.

Det finns vidare ett direktiv, Second Digital Payment Services (DPS2), som reglerar datatillgång till förmån för betaltjänstleverantörer.[23] I nämnda direktiv, artikel 36, ges tvingande tillgång för tillhandahållare av digitala betalningstjänster, till bankkontodata avseende individer som har blivit betaltjänstföretagets kunder.

Ett tredje exempel på sektorsspecifik tvingande datadelning finns i ett direktiv om data från smarta el- och gasmätare.[24] Ytterligare tre exempel på lagstadgad sektorspecifik datadelning är en förordning om elnätsdata,[25]

[22] En EU-strategi för data s. 7, från den 19.2.2020, COM(2020) 66 final.

[23] Direktiv 2015/2366 om betaltjänster.

[24] Direktiv 2019/944 om gemensamma regler för den inre marknaden för el, och direktiv 2009/73/EG om gemensamma regler för den inre marknaden för naturgas.

[25] Förordning (EU) 2017/1485 Riktlinjer för driften av elöverföringssystem, och Förordning (EU) 2015/703 om fastställandet av nätföreskrifter med regler för driftskompatibilitet och informationsutbyte.

ett direktiv om standarder för intelligenta transportsystem[26] och information som härrör från testning av kemikalier på ryggradsdjur[27].

EU-kommissionen tycks emellertid vara beredd att ta ytterliga och mer betydelsefulla steg mot lagreglerad tvingande datadelning. Ett uppmärksammat initiativ är förslaget från december 2020 till en Digital Market Act (DMA), som syftar till att skapa öppna och rättvisa digitala marknader.[28] Förordningen ska gälla för internets så kallade grindvakter, definierade i artikel 3, som de största internetplattformerna, bland annat karakteriserade genom att de har minst 45 miljoner månatligt aktiva användare. I dagsläget skulle Google, Apple, Facebook, Amazon, Microsoft och kanske några fler, omfattas.[29] EU-kommissionen uttalar att DMA syftar till att göra 'access to data … compulsory, where appropriate under fair, transparent, reasonable, proportionate and/or non-discriminatory conditions'.[30] I förslaget till DMA, artikel 6h och 6i, finns krav om att internetplattformarna ska ge företagsanvändare tillgång till data. Liknande regler finns sedan tidigare för personuppgifter i artikel 20 GDPR, men artikel 6 DMA gäller både personuppgifter och annan data.

> Artikel 6 DMA:
> h) tillhandahålla effektiv portabilitet för data som genereras genom en företagsanvändares eller slutanvändares verksamhet och ska i synnerhet tillhandahålla verktyg för slutanvändare för att underlätta utövandet av dataportabilitet, i enlighet med förordning (EU) 2016/679, bland annat genom tillhandahållande av kontinuerlig åtkomst i realtid,
>
> i) kostnadsfritt ge företagsanvändare, eller tredje parter som auktoriserats av en företagsanvändare, effektiv, högkvalitativ, kontinuerlig tillgång i realtid och användning av aggregerade eller icke-aggregerade data, som tillhandahålls eller genereras i samband med användningen av de relevanta centrala

[26] Direktiv 2010/40/EU ett ramverk för införande av intelligenta transportsystem på vägtransportområdet och för gränssnitt mot andra transportslag.

[27] Förordning (EG) nr 1907/2006 (Reach).

[28] Förslag från den 15.12.2020, COM(2020) 842 final, till EUROPAPARLAMENTETS OCH RÅDETS FÖRORDNING om öppna och rättvisa marknader inom den digitala sektorn (rättsakten om digitala marknader).

[29] L. Cabral, et al, The EU Digital Market – A Report from a Panel of Economic Experts s. 9, The European Commission science and knowledge service, 2021, tillgänglig på https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3783436.

[30] Communication from the Commission of 19 February 2020 to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions – A European strategy for data, COM(2020) 66 final, s. 13.

> plattformstjänsterna av dessa företagsanvändare och de slutanvändare som tar i anspråk i de produkter eller tjänster som dessa företagsanvändare tillhandahåller …

Artikel 6(h) garanterar effektiv dataportabilitet i realtid för företagsanvändare, förutsatt att slutanvändarna ger samtycke. Artikel 6(i) ska ge företagsanvändare fri och högkvalitativ tillgång till data i realtid som genereras av företagsanvändarens verksamhet på plattformen. Utöver att datadelningsskyldigheten i DMA utsträcks till alla slags data, och inte bara personuppgifter, omfattar således DMA möjligheter att begära hos internetplattformarna att få en kontinuerlig dataöverföring i realtid. En framtvingad datadelning av sådant slag syftar till att minska de fördelar som plattformarna får genom data som genererats genom aktivitet på plattformen, vilket innebär att andra än plattformen kan utveckla nya tjänster baserat på plattformsdata. Viktigt att notera är att den tvingande datadelningen från de största internetplattformarna bara gäller företagsanvändarens egen data som finns på plattformen; det finns ingen rätt att få tillgång till andra företags data som finns hos plattformen. Internetplattformen kommer således ensam att kunna kombinera data från olika deltagande företags verksamhet. Alternativet, att varje företag skulle kunna få tillgång till alla på plattformen deltagande företags data, vore emellertid mycket långtgående och kanske riskera hela plattformens affärsverksamhet.[31]

Som komplement till sektorspecifika regler om tvingande delning, har EU-kommissionen försökt att på olika sätt undanröja praktiska hinder mot frivillig avtalsbaserad spridning av data. I förordningen om det fria flödet av andra data än personuppgifter, finns ett förbud mot att EU-medlemsstater uppställer krav på (nationell) datalokalisering (EU-länder måste godta datalagring i andra medlemsstater) och en skyldighet för marknadsaktörer att införa uppförandekoder som underlättar dataportering vid uppsägning av avtal.[32] Senast 29 november 2022 ska

[31] L. Cabral, et al, The EU Digital Market – A Report from a Panel of Economic Experts s. 22, The European Commission science and knowledge service, 2021, tillgänglig på https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3783436.

[32] ”För att dra nytta av den konkurrensutsatta miljön fullt ut bör professionella användare kunna göra välinformerade val och på ett enkelt sätt jämföra enskilda delar av olika erbjudanden om databehandlingstjänster på den inre marknaden, bland annat när det gäller avtalsvillkoren för dataportering vid uppsägning av avtal. Den detaljerade informationen och de operativa kraven för dataportering bör, i syfte att anpassa dem till marknadens

Kommissionen utvärdera om nämnda uppförandekoder har etablerats och genomförts effektivt, samt huruvida tjänsteleverantörerna verkligen tillhandahåller information som underlättar dataportering.

4 Artikel 102 Funktionsfördraget och tvingande datadelning

Ett företags vägran att dela data kan i det enskilda fallet utgöra missbruk av en dominerande ställning, enligt artikel 102 Funktionsfördraget. Vägran att dela data har bedömts som ett missbruk av en dominerande ställning i det så kallade Magill-målet[33], som gällde tre tv-bolags vägran att dela information om framtida tv-tablåer med en fristående utgivare av en tv-guide. Informationen behövdes för att Magill skulle kunna introducera en tv-guide som veckovis innehöll samtliga tre irländska tv-bolags program, vilket var en förbättring jämfört med tv-bolagens egna guider som bara innehöll respektive kanals program. EU-domstolen fann att det var ett missbruk av en dominerande ställning att neka en rätt att använda programtablåerna, när Magill skulle lansera en ny produkt som det existerade en efterfråga för.

En parallell situation skulle vara att ett företag vill lansera en ny AI-baserad tjänst och behöver få tillgång till andra företags unika realtidsdata. I ett sådant fall skulle utvecklaren av den AI-baserade tjänsten för-

innovationspotential och med beaktande av den erfarenhet och sakkunskap som finns hos tjänsteleverantörer och professionella användare av databehandlingstjänster, fastställas av marknadsaktörerna genom självreglering i form av uppförandekoder på unionsnivå som skulle kunna inbegripa standardavtalsvillkor, vilket bör uppmuntras, underlättas och övervakas av kommissionen.” Preambel 30 i Förordning (EU) 2018/1807.

”För att uppnå ändamålsenlighet och underlätta byte av tjänsteleverantör och dataportering bör uppförandekoderna vara heltäckande och inbegripa åtminstone de huvudaspekter som är viktiga under dataporteringsprocessen, såsom de processer som används för, och platsen för, backup av data, tillgängliga dataformat och support, erforderlig it-konfiguration och minsta nätverksbandbredd, den tid som krävs innan porteringsprocessen inleds och den tid under vilken data kommer att förbli tillgängliga för portering samt garantier för tillgång till data om tjänsteleverantören går i konkurs. Uppförandekoderna bör även klargöra att inlåsning till en leverantör inte är en godtagbar affärspraxis samt föreskriva tillitsfrämjande teknik, och de bör uppdateras regelbundet för att hålla jämna steg med den tekniska utvecklingen.” Preambel 31 i Förordning (EU) 2018/1807.

[33] Förenade målen C-241/91 och C-242/91, RTE och ITP mot Kommission, [1995] ECR I-743 = ECLI:EU:C:1995:98.

modligen, i likhet med Magill, behöva visa att informationen inte kunde erhållas på annat sätt.[34] Det behöver förmodligen också göras troligt att tjänsten är efterfrågad och inte skulle kunna tillhandahållas utan tredje mans hemlighållna data, dvs. att det är omöjligt för tjänsteutvecklaren att själv samla in data eller att erhålla den från annat håll. Man kan i det sammanhanget fråga sig om det har någon betydelse att det brukar vara svårbedömt om data tekniskt sett kommer att möjliggöra en ny tjänst. I Magill-målet var det tydligt att en tv-guide kunde produceras om bara programinformationen erhölls; osäkerheten är flerfalt större vid industriell data och det blir i motsvarande mån mer tveksamt att framtvinga datadelning för vad som kanske är utomståendes ganska spekulativa försök att utveckla tjänsteinnovationer.

Vidare finns Microsoft-målet,[35] som gällde en skyldighet för företaget att dela den information som konkurrenter behövde för att kunna tillverka serveroperativsystem kompatibla med Microsofts Windows. Kravet i Magill-målet på att en ny produkt skulle tas fram av konkurrenten för att leveransvägran skulle vara missbruk, vidareutvecklades i målet.[36] Tribunalen/Förstainstansrätten fann att regeln om att leveransvägran ska vara ett hinder mot en ny produkt – nyproduktkriteriet – knappast gäller när företaget har en mycket stark ställning, en super-dominans. Vid särskilt stark marknadsmakt kan man ta hänsyn till att det i ett längre perspektiv är sannolikt att nya produkter från tredje män hindras om inte nödvändig interoperabilitetsdata delas. Det finns emellertid ännu ingen ytterligare europeisk rättspraxis som bekräftar att det inte ska krävas en avsikt att ta fram en ny produkt.[37] Frågan är kanske inte avgörande vid utveckling av AI-tjänster. Det är förmodligen så att den som försöker

[34] J. Drexl, Connected devices – An unfair competition law approach to data access rights of users, DOI:10.5771/9783748924999-477, publicerad i boken Data Access, Consumer Interests and Public Welfare, s. 477 (506), och tillgänglig på https://www.researchgate.net/publication/350522751_Connected_devices_-_An_unfair_competition_law_approach_to_data_access_rights_of_users.

[35] T-201/04 Microsoft [2007] ECR II-3601 = ECLI:EU:T:2007:289, stycke 334.

[36] Se närmare i B. Mäihäniemi, Competition Law and Big Data – Imposing Access to Information in Digital Markets s. 184 (2020).

[37] J. Drexl, Connected devices – An unfair competition law approach to data access rights of users, DOI:10.5771/9783748924999-477, publicerad i boken Data Access, Consumer Interests and Public Welfare, s. 510, och tillgänglig på https://www.researchgate.net/publication/350522751_Connected_devices_-_An_unfair_competition_law_approach_to_data_access_rights_of_users.

framtvinga datadelning för att ta fram en AI-tjänst avser att lansera något som inte redan finns på marknaden; nyproduktkriteriet är uppfyllt.

Sammanfattningsvis finns vissa, men mycket begränsade möjligheter, att kräva datadelning med stöd av artikel 102 FEUF, för den som avser att träna en AI-baserad algoritm med hjälp av annans data/företagshemligheter. Det ska vara omöjligt att få tillgång till data på annat sätt. Vidare måste innovatören göra sannolikt att det kommer att tas fram en efterfrågad och ny tjänst. Missbruk av dominerande ställning genom en vägran att dela data kan förekomma inom alla branscher, men kriterierna är så högt ställda att förmodligen mycket få potentiella innovatörer inom AI-tjänster kan räkna med det. Det är en utdragen och osäker process som behöver inledas i en situation där snabbhet är avgörande. Möjligheten finns bara vid klara möjligheter till väsentlig teknisk utveckling som av någon anledning hindras av en dominant aktör.[38]

5 Slutsatser

Allt pekar mot en framtid där industriell data växer i betydelse, eftersom den har potential att optimera oräkneliga processer i stora delar av samhället. EU har slagit in på vägen att inte skapa en ny immaterialrätt för sådana data, utan att använda reglerna för företagshemligheter. Fokus har skiftat mot att fram för allt underlätta, men att i exceptionella fall framtvinga, datadelning. Det senare sker genom fall av sektorspecifik lagstiftning kompletterat av artikeln 102 Funktionsfördraget (dominerande företag hindrar att en ny väsentlig produkt eller tjänst kan introduceras).

Det är rimligt att en möjlighet till tvingande datadelning undantagsvis står till förfogande, men sådana drastiska ingripanden måste användas med stor försiktighet.[39] Vid en frivillig licensiering av industriell data vet man att värdet av licenstagarens nyttjande överstiger värdet som en exklusiv rätt till samma data hade haft för licensgivaren. Vid licensiering har genomförts en jämförelse mellan parternas framtida användningsmöjligheter. Efter en framtvingad datadelning finns ingen sådan säkerhet om att värdet ökar. Samtidigt är det klart att transaktionskostnaderna

[38] D. Eklöf, Upphovsrätt i konkurrens s. 432 (2005).

[39] J. Drexl, Connected devices – An unfair competition law approach to data access rights of users, DOI:10.5771/9783748924999-477, publicerad i boken Data Access, Consumer Interests and Public Welfare, s. 480, och tillgänglig på https://www.researchgate.net/publication/350522751_Connected_devices_-_An_unfair_competition_law_approach_to_data_access_rights_of_users.

vid industriell data är betydande och man kan således inte utesluta att framtvingad datadelning leder till nya tjänster som ger en samhällelig värdeökning, jämfört med företags helt exklusiva position i fråga om egeninsamlad data. I särskilda fall kan det således vara motiverat att lagstiftaren eller domstolen finner att licenstagarens nyttjande är mer värt än innehavarens exklusiva tillgång.

Ett särskilt skäl till återhållsamhet med tvingande datadelning är områdets tekniska komplexitet, snabbt växande datatillgång och affärsmodeller som är i konstant utveckling.[40] Det är en påtaglig risk att eventuell lagstiftad eller domstolsbeslutad tvingande datadelning blir olämplig eller åtminstone snabbt föråldrad (om den var riktig vid bedömningsögonblicket). Ytterligare skäl till återhållsamhet med tvingande delning, är att information från ett annat företags sensorer i sig sällan är tillräckligt; det krävs kontextuell information om hur data har insamlats, under vilken tidsperiod, m.m. En tillförlitlig och fungerande AI-tjänst måste ha kvalitativ och användbar data. Normalt förutsätter detta frivillighet mellan parterna. En innehavare som inte frivilligt delar data kommer att vara en tveksam dataleverantör och regler om datakvalitet är svåra att hantera vid tvingande datadelning. Man har försökt att lösa detta i artikel 6(i) DMA genom att stipulera ”effektiv, högkvalitativ, kontinuerlig tillgång i realtid” beträffande data som ett företags verksamhet på plattformen genererar. Men det är osäkert i vilken grad som kvalificeringen ”högkvalitativ” data verkligen ger mottagaren den kontext och kvalitet som behövs. Det sagda indikerar en svårighet med DMA-förslagets tvingande datadelning för stora internetplattformar.

Sammanfattningsvis är det mycket som krävs innan tvingande datadelning framstår som ett alternativ. Ett tydligt och bestående marknadsmisslyckande ska vara för handen, alltså att värdehöjande avtal kring data under en längre tid har visat sig vara omöjliga. Med det sagt, anser jag att den europeiska utvecklingen kring industriell data hitintills förefaller att vara på rätt spår. Man förlitar sig på skyddet för företagshemligheter och möjligheterna till frivilliga avtal, men undantagsvis ges efterföljande datainnovatörer en hjälpande hand. Det kan vara betaltjänstutvecklare, innovatörer som vill använda elnätsdata eller kanske skapare av nya hemsidor som på ett innovativt sätt sorterar veckans kommande medie-

[40] J. Drexl, Designing Competitive Markets for Industrial Data s. 9, Max Planck Institute for Innovation and Competition Research Paper No. 16-13, tillgängligt på https://ssrn.com/abstract=2862975.

innehåll. En första innovatör brister ibland när det gäller förmågan att insé och viljan att ta till vara, tekniska vidareutvecklingar.[41] Tredje man som ser möjligheter kan i ett sådant läge ha svårt att kliva fram och få till stånd de avtal om data som behövs. I en alltmer digital ekonomi med potentiellt många, men svårförutsebara och immaterialrättsligt oskyddade nya AI-baserade tjänster, kan problemet med tillgång till data som inte är personuppgifter, bli ett märkbart utvecklingshinder, särskilt i tidiga faser av nyttiggörandet, då potentialen hos industriell data utforskas.

[41] Empiriska exempel finns i R. P. Merges, R. R. Nelson, On the Complex Economics of Patent Scope, 90 Columbia Law Review s. 839.

Markku Suksi

Lagbundenhetskravet vid automatiserat beslutsfattande i myndighetsverksamhet enligt finsk rätt

1 Inledning

Lagbundenhetsprincipen betyder för myndighetsverksamheten att all utövning av offentlig makt ska bygga på lag och att lag ska noggrant iakttas i all offentlig verksamhet. Lagbundenhetsprincipen, som vanligtvis antar den från tysk rätt bekanta principen om rättsstatlighet, *Rechtsstaatlichkeit*, utgör en grundpelare för all myndighetsverksamhet både i Finland och Sverige och i en mängd andra länder inom och utom Europa. I Finland ingår utgångspunkten för lagbundenhetsprincipen i grundlagens 2 § 3 mom. och i Sverige i regeringsformens 1 § 3 mom., men därtill finns andra bestämmelser i bägge ländernas grundlagar som hör ihop med lagbundenhetsprincipen och rättsstatligheten.

Även om den lagstiftning som gäller myndighetsverksamhet i princip är teknologineutral, är denna lagstiftning på det stora hela stiftad utifrån föreställningen om att myndigheternas beslut fattas av fysiska tjänsteinnehavare. Kravet på lagbundenhet vid myndighetsutövningen har därmed främst gällt offentlig beslutsverksamhet genom enskilda tjänstemän och genom kollegiala organ mot bakgrunden av en princip om tjänstemannaförvaltning. Det finns därtill exempel på lagstiftning som för beslutsverksamhetens del ger uttryckligen för handen att besluten ska fattas av fysiska tjänsteinnehavare eller förtroendevalda, såsom den finska kommunallagens (410/2015) 91 § 1 mom., första satsen.[1] Den svenska kom-

[1] ”Fullmäktige kan i förvaltningsstadgan delegera beslutanderätt till kommunens övriga organ samt till förtroendevalda och tjänsteinnehavare.”

munallagens (2017:725) delegeringsbestämmelser är likartade beträffande förutsättningen om att delegera beslutsmakt till nämnder, direktioner och tjänsteinnehavare, dvs. till fysiska kommunala beslutsfattare, även om den svenska förvaltningslagen (2017:900) innehåller i 28 § en tillåtande bestämmelse om användning av automatiserat beslutsfattande. Någon sådan grundläggande bestämmelse som skulle motsvara 28 § finns inte – åtminstone ännu – i finsk lagstiftning.

I takt med ökad automatisering av beslutsfattandet vid myndigheter har man börjat fråga sig efter lagstöd för algoritmiskt beslutsfattande, dvs. vid beslutsfattande som är automatiserat och sker utan att någon fysisk tjänsteinnehavare eller något organ med fysiska personer deltar i beslutsfattandet. Härvid finns givetvis ingen orsak att tänka sig att den statsrättsliga lagbundenhetsprincipen inte skulle gälla för automatiserat beslutsfattande, dvs. att ett offentligt samfund kunde åsidosätta lagbundenhetsprincipen genom att överflytta beslutsfattandet från mänskliga beslutsfattare till en algoritmbaserad beslutsfattare. Frågorna som inställer sig är hur lagbundenhetsprincipen ska gälla för automatiserat beslutsfattande och hur det automatiserade beslutsfattandet ska regleras för att tillfredsställa de krav som lagbundenhetsprincipen uppställer? Automatiserat beslutsfattande ska givetvis äga rum inom ramen för det konstitutionella kravet på lagbundenhet, men på vilket sätt ska lagbundenhetsprincipen konkretiseras i sådant beslutsfattande som är algoritmbaserat?

Lagbundenhetsprincipens och rättsstatlighetens betydelse inom förvaltningsrätten är ett tema som behandlats i många sammanhang. Enligt Alexander Pezcenik förutsätter rättsstaten i sin ideala typ ”att den offentliga makten utövas inom en rättslig ram” som gör maktutövningen förutsebar med stöd av rättsregler.[2] Enligt Pezcenik förutsätter rättsstaten att även maktutövningens innehåll regleras av rättsnormerna, vilket betyder att ”medborgarna måste kunna veta allt väsentligt om *hur* makten kommer att utövas, inte endast vem som kommer att göra det”.[3] Därmed blir maktutövningen inte godtycklig. En färsk sammanfattning av rättsstatlighet ingår i Ida Asplunds doktorsavhandling år 2021, där legaliteten inom förvaltningen anses bestå av tre olika dimensioner hos rättsstatlighet, nämligen formell rättsstatlighet, processuell rättsstatlig-

[2] Alexander Pezcenik, *Vad är rätt? Om demokrati, rättssäkerhet, etik och juridisk argumentation*. Stockholm: Norstedts Juridik, 1995, s. 50.

[3] Ibidem.

het och materiell rättsstatlighet.[4] Med formell rättsstatlighet kan avses att beslutsfattandet grundar sig på lag, med processuell rättsstatlighet att olika i lagstiftning etablerade förfarandemässiga säkringsåtgärder kringgärdar beslutsfattandet vid myndigheter och med materiell rättsstatlighet att de innehållsmässiga förutsättningarna för förvaltningsbeslutet är fastslagna i lagstiftningen. Avsikten med denna artikel är inte att granska svensk praxis kring automatiserat beslutsfattande, utan att analysera finsk praxis från riksdagens grundlagsutskott, riksdagens justitieombudsman och statsrådets justitiekansler med avseende på automatiserat beslutsfattande och de krav som ställs på detta utifrån den finska rättsordningen genom att försöka placera de juridiska uttalandena i ovannämnd praxis i de tre kategorierna av rättsstatlighet; formell, processuell och materiell rättsstatlighet.

2 Lagövervakares observationer om lagbundenhetsprincipen

Biträdande justitieombudsmannen Maija Sakslin granskade två av Skatteförvaltningens automatiserade förfaranden i anslutning till skatter vars deklarationer är självinitierade av den skattskyldige.[5] I det fall att Skatteförvaltningen emottar en felaktig eller logiskt sett bristfällig deklaration, översänder det automatiserade förfarandet en uppmaning eller ett brev med utredningsbegäran. I ett ärende fanns en motstridighet beträffande den skattskyldiges skatteslag som inte avlägsnades till följd av olika automatiskt skickade begäranden om utredning. Denna motstridighet ledde till ett beskattningsbeslut som fattades genom automatiserat förfarande och som inbegrep en skattehöjning om 20 procent jämte dröjsmålspåföljd. Den skattskyldige påfördes en skatt som han redan betalat, något som i ett senare skede undanröjdes till följd av rättelseyrkande. I ett annat ärende hade Skatteförvaltningens automatiserade förfarande skickat

[4] Ida Asplund, *Den enskildes rättssäkerhet i individnära tillsyn*. Umeå: Umeå universitet, 2021, ss. 81, 102.

[5] Avgörande av biträdande justitieombudsman Maija Sakslin, 20.11.2019, EOAK/3379/2018: Skatteförvaltningens automatiserade beslutsförfarande uppfyller inte grundlagens krav, https://www.oikeusasiamies.fi/r/fi/ratkaisut/-/eoar/3379/2018 (besökt den 9 augusti 2021). För ett sammandrag på svenska, se Riksdagens justitieombudsmans berättelse år 2019, ss. 351–352, https://www.riksdagen.fi/SV/vaski/JulkaisuMetatieto/Documents/B_15+2020.pdf.

till 11 000 primärproducenter brev med grundlösa uppmaningar i vilka de skattskyldiga hotades med beskattning enligt prövning och skattehöjningar.[6] Enligt Skatteförvaltningens utredningar föranleddes de aktuella skattskyldiga inga skattepåföljder av det felaktiga förfarandet.

Oavsett de företagna korrigeringarna ansåg biträdande justitieombudsmannen i det första fallet med meddelandet av Skatteförvaltningens allmänna telefonnummer för upplysningar i beskattningsfrågor, där förfrågningarna besvarades av personer som inte hade varit med om beslutsprocessen för de aktuella ärendena, att det automatiserade beskattningsförfarandet skapar oklarhet och osäkerhet om hur den skattskyldige kan få saklig och sakkunnig personlig service och rådgivning och upplysningar om grunderna för beslutsfattandet. Oavsett de uteblivna skattepåföljderna ansåg biträdande JO i det andra ärendet att den skattskyldiges rättsskydd förefaller att äventyras om de utredningar som den skattskyldige lämnat med anledning av automatiserade uppmaningar om tilläggsutredningar inte granskas på ett sakligt sätt, med den konsekvensen att det bristande förfarandet leder till ett automatiserat beskattningsbeslut med en skattehöjning om 20 procent jämte dröjsmålspåföljd.

Biträdande JO konstaterade att beskattning utgör betydande utövning av offentlig makt som inbegriper självständigt bruk av prövningsrätt och rätt att på ett betydande sätt ingripa i individens rättigheter och skyldigheter. Hon fortsätter med att konstatera att det i grundlagens 2 § stadgas bland annat om principen om rättsstatlighet, där 3 momentet förutsätter att offentlig makt bör grunda sig på lag och att lagen bör noggrant följas

[6] Skatteförvaltningens användning av automatiserat beslutsfattande är omfattande, såsom framgår av Riksdagens justitieombudsmans berättelse år 2019, s. 352: ”Skatteförvaltningens förfarande för påförande av skatt enligt uppskattning, som används för skatter som betalas på eget initiativ, är automatiserat. Skatteförvaltningen sänder årligen ut cirka 300 000 uppmaningsbrev med anledning av uteblivna anmälningar och meddelar drygt 112 000 skattebeslut baserade på uppskattning via automatiserad handläggning. I dessa fall har informationssystemet hanterat alla handläggnings- och beslutsfaser utan att en fysisk person har varit in-blandad. I den automatiserade beskattningen baserad på uppskattning har Skatteförvaltningen också påfört skatteförhöjningar på 25 procent av det uppskattade skattebeloppet. Likaså meddelas automatiserade beskattningsbeslut baserade på uppskattning i inkomstbeskattningen för samfund, och skatteförhöjningar på fem procent av det uppskattade skattebeloppet påförs. Även skattekontroll och skatteuppbörd, förutom de fall som styrs till enskild behandling, sker automatiserat enligt förinställda systemkonfigurationer. En skattskyldigs skatteärende kan med andra ord genomgå hela handläggningsprocessen med hörande, beslut och uppbörd enligt ett automatiserat förfarande, utan att en enda fysisk person har deltagit i handläggningen.”

i all offentlig verksamhet. Hon konstaterar också att grundlagens bestämmelser om principen om förvaltningens lagbundenhet i kombination med grundlagens bestämmelser om tjänstemannaansvar ger uttryck för principen om tjänstemannaförvaltning. Det centrala i hennes budskap är att de bestämmelser som finns i skattelagstiftningen om automation avser sättet att utreda, inte själva beslutsfattandet: ”Författningsgrunden för Skatteförvaltningens automatiserade förvaltnings- och beslutsförfarande uppfyller inte grundlagens krav.”[7] Det syns att skattelagstiftningen saknade en i grundlagens 2 § 3 mom. avsedd grund för befogenheten att fatta automatiserade beslut.

Även om merparten av de bestämmelser i grundlagen och i vanlig lag som biträdande JO:s avgörande hänvisade till gällde annat än lagbundenhetsprincipen, är slutsatsen i biträdande JO:s avgörande nära förbunden med rättsstatligheten: eftersom Skatteförvaltningens automatiserade beskattnings- och beslutsförfarande inte baserar sig på en saklig och noggrann reglering i lag som skulle beakta det sakliga förverkligandet av god förvaltning, rättsskydd och tjänsteansvar, anser hon att det automatiserade beskattnings- och beslutsförfarandet är lagstridigt. Det kan vara möjligt att argumentera att här antyds något om både formell och processuell lagbundenhet i och med att biträdande JO hänvisar, utöver grundlagens 2 § 3 mom., även till grundlagens 21 § om rätten till god

[7] Riksdagens justitieombudsmans berättelse år 2019, s. 351. Samma år avgjorde biträdande JO Sakslin två ytterligare fall och fann dem lagstridiga, dock främst med hänvisning till principer om god förvaltning och bestämmelser i förvaltningslagen. I avgörandet 2216/2018 ansåg biträdande JO att i Skatteförvaltningens utredning eller utlåtande till den skattskyldige ”gav man inte den viktiga informationen om att betalningsarrangemanget behandlas i det automatiserade förfarandet om skatten inte överstiger 10 000 euro. Enligt BJO var det här en viktig och väsentlig detalj när ärendet prövades, men man var tvungna att skilt fråga efter informationen.” I avgörandet 2898/2018 ansåg biträdande JO Sakslin följande: ”Av beslutet som fattats i det automatiserade förfarandet framgick inte för den klagande på vilket sätt man i beslutsfattandet beaktat de rättsprinciper som ingår i grunderna för god förvaltning och som styr förvaltningen och förvaltningens beslutsfattande. Det framgick inte heller av beslutet om man i beslutsfattandet överhuvudtaget prövat de grunder och motiveringar som den klagande framställt, enligt vilka det enligt den klagande fanns ett motiverat skäl i ärendet att lämna in anmälan på ett sätt som avviker från Skatteförvaltningens beslut. Av beslutet framgick inte heller varför de framställda motiveringarna inte ansågs tillräckliga. Av beslutet framgick således inte så som förvaltningslagen förutsätter vilka omständigheter och utredningar som inverkat på avgörandet och inte heller namnet på och kontaktuppgifterna till den person som parten vid behov kan kontakta för att få mera information om beslutet.” Se Riksdagens justitieombudsmans berättelse år 2019, ss. 352–353.

förvaltning och rättsskydd och till grundlagens 118 § om tjänsteinnehavarnas straff- och skadeståndsrättsliga ansvar. En central observation av biträdande JO är att automatiserat beslutsfattande har utvecklats inom Skatteförvaltningen utan att tillräcklig utredning och bedömning av behovet av att i lag stifta om förfarandet på ett sådant sätt som skulle ha beaktat synpunkter som riksdagens grundlagsutskott framfört om sådana frågor som var öppna eller outredda. ”Skatteförvaltningens automatiserade förvaltnings- och beslutsförfarande måste regleras genom precisa och exakta lagbestämmelser. Denna reglering borde bl.a. definiera hur man väljer de ärenden som ska skötas genom det automatiserade beslutsförfarandet och hur offentligheten för förfarandets algoritmer tillgodoses. För att algoritmerna ska omfattas av tillbörlig offentlighet i en form som gemene man förstår måste det definieras exakt och precist i lagen vad som avses med en algoritm i ett automatiserat beslutsförfarande.”[8]

Statsrådets justitiekansler Tuomas Pöysti gav år 2021 ett avgörande om Folkpensionsanstaltens (FPA) automatiserade beslutsfattande[9] där han framför liknande tankegångar. I ett avsnitt om den rättsliga grunden för automatiserat beslutsfattande konstaterar JK Pöysti, att den är otillräcklig. FPA har till uppgift att verkställa lagstiftning inom området för socialskydd genom vilken det allmänna förverkligar den rätt till socialskydd som ingår i grundlagens 19 §. I de förvaltningsbeslut som fattas i denna verksamhet bör FPA efterleva olika förfaranderegler om god förvaltning och behandling av förvaltningsärenden och trygga kundernas rättigheter. En del av dessa beslut fattas genom automatiserad behandling av kundernas personuppgifter som hänför dylik behandling till området för allmänna dataskyddsförordningens artikel 22. Det saknas emellertid sådan rättslig grund för automatiserade beslut vid FPA som grundlagens 80 § och dataskyddsförordningens artikel 22(2) förutsätter.

Avsaknaden av tillräckligt entydig rättsgrund för automatiserat beslutsfattande är något som FPA medger i sin utredning till JK. Därtill efterlyser FPA särlagstiftning om automatiserat beslutsfattande i tillägg till allmän lagstiftning om automatiserat beslutsfattande. Enligt JK är det vid automatiserat beslutsfattande i sista hand fråga om förverkligan-

[8] Riksdagens justitieombudsmans berättelse år 2019, ss. 351–352.

[9] JK:s avgörande 20.4.2021, OKV/131/70/2020, https://www.okv.fi/media/filer_public/05/29/0529702c-beaf-4f9e-8111-377ebdd0a2ba/okv_131_70_2020.pdf (besökt den 10 augusti 2021). För en svenskspråkig sammanfattning, se https://www.okv.fi/sv/meddelanden/563/justitiekansler-tuomas-poysti-det-finns-behov-att-foreskriva-om-fpas-automatiserade-beslutsfattande-i-lag/ (besökt den 10 augusti 2021).

det av de rättigheter som förvaltningens kunder har, men JK misstänker inte att FPA skulle ha misslyckats med att trygga kundernas rättigheter och rättsskydd. Det som JK konstaterar är att läget är oreglerat inte enbart beträffande den allmänna lagstiftning som gäller för automatiserat beslutsfattande,[10] utan också beträffande sådan speciallagstiftning som eventuellt behövs. JK framför att fungerande och tillräcklig lagstiftning är ett sätt att försäkra förverkligandet av god förvaltning och kundernas rättsskydd vid automatiserat beslutsfattande och de beslutssystem som aktualiseras. Därför efterlyser JK en korrigering av det oreglerade rättsläget och föreslår att regleringsbehovet beträffande FPA:s automatiserade beslutsfattande utreds på ett helhetsmässigt sätt uttryckligen med tanke på sådan särlagstiftning om automatiserat beslutsfattande som FPA:s verksamhet föranleder.

I och med att det för FPA:s automatiserade beslutsfattande inte finns någon sådan rättslig grund som förutsätts i den allmänna dataskyddsförordningen eller i grundlagen kan man mot bakgrunden av JK-avgörandet konstatera att detta avgörande, utan att i någon större utsträckning än nämna lagbundenhetsprincipen i grundlagens 2 § 3 mom. (förutom att framhålla att bestämmelsen är knuten till grundlagens 118 § om tjänsteansvar, som kompletterar lagbundenhetsprincipen), hänsyftar till denna grundlagsbestämmelse om lagbundenhet. Med andra ord, om den rättsliga grund som förutsätts av grundlagens 80 § och allmänna dataskyddsförordningens artikel 22 inte finns, uppkommer åtminstone ett indirekt problem beträffande lagbundenhetsprincipen. JK är medveten om att beredning av allmän lagstiftning om automatiserat beslutsfattande pågår under 2021 vid justitieministeriet och han anser att sådan lagstiftning, efter att den godkänts, kommer att lösa vissa frågor som gäller automatiserat beslutsfattande. Det här kan tolkas som en hänvisning till processuell rättsstatlighet. JK är emellertid inte övertygad om att allmän

[10] I avgörandet av ställföreträdaren för biträdande justitiekanslern Petri Martikainen av den 4 september 2019, OKV/868/1/2018, konstateras emellertid att automatisering av beslutsfattandet inte berättigar till avsteg från de krav som förvaltningslagen ställer. Enligt förvaltningslagens 44 § 1 mom. angående beslutets innehåll bör av ett skriftligt beslut tydligt framgå namn och kontaktuppgifter för den person av vilken en part vid behov kan begära ytterligare uppgifter om beslutet. FPA:s beslut som delgivits klaganden uppfyllde inte dessa krav. Bestämmelsens syfte och de krav som god förvaltning och särskilt serviceprincipen ställer skulle bäst uppfyllas av ett förfarande, där det av beslutet framgår namnet och kontaktuppgifterna på den förmånshandläggare som berett ärendet och därtill kontaktuppgifter till förmånsspecifika kundrådgivare.

lagstiftning ensam räcker till för att hantera de problem som är förknippade med automatisering av beslutsfattandet vid FPA. Därför föreslår han att behovet av speciallagstiftning som gäller FPA:s automatiserade beslutsfattande utreds. I detta senare avseende förefaller JK tänka i termer av materiell rättsstatlighet, dvs. att den materiella lagstiftning som FPA tillämpar behöver förses med särskilda bestämmelser om automatiserat beslutsfattande. Det skulle i sin tur betyda att materiell lagstiftning bör innehålla bemyndiganden om ibruktagande av automatiserat beslutsfattande för sådana beslutskategorier där dylikt förfarande är tänkbart och därtill kanske innehållsmässiga utgångspunkter för automatiserat beslutsfattande i förmånslagstiftningen.

Det är inte helt okomplicerat att placera de två lagövervakarnas uttalanden i de tre kategorierna av rättsstatlighet, men det syns att de uppvisar något olika profiler: biträdande justitieombudsmannens uttalanden länkar in på formell och processuell rättsstatlighet, medan justitiekanslerns uttalanden länkar in på processuell och materiell rättsstatlighet. Det som är gemensamt är att de har konstaterat lagstridigheter av olika slag vid användningen av automatiserat beslutsfattande och att de efterlyser lagstiftning om automatiserat beslutsfattande. Båda lagövervakarna är medvetna om grundlagsutskottets ställningstaganden med avseende på automatiserat beslutsfattande (se nedan), vilket betyder att lagövervakarnas avgöranden är delvis parallella i förhållande till grundlagsutskottets uttalanden i anslutning till lagstiftningsförfarande. Avgörandena har således tillkommit i en aktiv fas av arbetet med att försöka greppa den juridiska problematiken med automatiserat beslutsfattande, vilket betyder att graden av angelägenhet hos avgörandena är hög.

3 Grundlagsutskottets observationer om lagbundenhetsprincipen

3.1 Inledande steg om automatiserat beslutsfattande

Såsom framgått av det ovan anförda så har den finska lagstiftaren och särskilt riksdagens grundlagsutskott som den auktoritativa uttolkaren av grundlagens bestämmelser dryftat frågor i anslutning till automatiserat beslutsfattande och rättsstatlighet. Ett tidigt utlåtande av grundlagsutskottet utgörs av GrUU 51/2016 rd som avgavs med anledning av regeringens proposition till riksdagen med förslag till lag om försök med basinkomst och lag om temporär ändring av 92 § i inkomstskattelagen och 17 § i lagen om förskottsuppbörd. Grundlagsutskottet fäste med anledning av kravet på bestämmelser i lag även uppmärksamhet vid urvalsförfarandet och dess transparens. Enligt 5 § 1 mom. i det första lagförslaget tar Folkpensionsanstalten genom ett slumpmässigt urval ut 2 000 personer till vilka det betalas basinkomst, något som i sig föranledde utskottet att framföra anmärkningar och ändringsförslag mot bakgrunden av grundlagens 6 § om likställdhet och icke-diskriminering. ”Utskottet anser att försöksgruppen inte kan väljas ut enbart med hjälp av en programkod, utan både kravet på bestämmelser i lag och kraven på att de ska vara exakta och noga avgränsade förutsätter bestämmelser i lag om grunderna för urvalet, t.ex. enligt krav som motsvarar karaktäriseringarna i motiven till lagstiftningsordning. Bestämmelserna måste kompletteras till denna del.”

Därtill framförde grundlagsutskottet att lagen bör även innehålla uttryckliga bestämmelser om offentliggörandet av programkoden och dess offentlighet. Ingen hänvisning till grundlagens 2 § 3 mom. ingår i utlåtandet, så man får anta utskottet med kravet på bestämmelser i lag avser det lagförbehåll som ingår i grundlagens 6 § 2 mom. om icke-diskriminering. En sådan hänvisning kan emellertid åtminstone indirekt förverkliga rättsstatsprincipen, kanske främst i den processuella meningen. Social- och hälsovårdsutskottet införde det av grundlagsutskottet föreslagna innehållet i sitt betänkande och riksdagen antog lagen om försök med basinkomst (1528/2016) 5 § 1 mom. med bl.a. följande innehåll: ”Det slumpmässiga urvalet görs på så sätt att var och en som hör till målgruppen har lika möjlighet att bli vald till försöksgruppen. Folkpensionsanstalten ska innan försöket inleds publicera den programkod som ska användas för samplingen.” Det är inte här fråga om egentliga förvaltningsbeslut som avgör enskild tillkommande fördel, rättighet eller skyldighet, dvs. vad individen kommer att få i basinkomst, men sammanställandet av försöksgruppen genom algoritmiskt förfarande kan ändå anses vara ett automatiserat beslut.

3.2 Tilltagande konstitutionella krav

Automatiserade beslut vid myndigheter och sådana privata försäkringsbolag som handhar lagstadgade försäkringar var föremål för regeringens proposition RP 52/2018 rd med förslag till ändring av socialtrygghets- och försäkringslagstiftningen med anledning av EU:s allmänna dataskyddsförordning. Enligt utlåtandet GrUU 78/2018 rd skulle det föreslagna beslutsfattandet få bygga på automatisk behandling ”bara om det med beaktande av det behandlade ärendets art och omfattning samt kraven enligt denna lag och kraven på god förvaltning är möjligt att meddela ett automatiserat beslut”. Den föreslagna lagen skulle dock inte ange detaljerat i vilka situationer eller inom vilka ärendegrupper behandlingen kan ske helt automatiskt, men på den punkten ansåg grundlagsutskottet att regleringen måste preciseras: ”I regleringen om automatiserat beslutsfattande måste det anges mer exakt än i förslaget på vilka grunder ärenden kan avgöras genom automatiserat beslutsfattande.”

Detta uttalande går i riktning mot materiell rättsstatlighet, men utskottet förutsatte också ändringar som understryker processuell rättsstatlighet genom att understryka vikten av hörande av part och andra principer som hör till god förvaltning enligt grundlagens 21 § och förvaltningslagen. Även av dessa orsaker behövde bestämmelserna om automatiserat beslutsfattande preciseras. Utskottet hänvisar emellertid också till principen om att utövning av offentlig makt bör basera sig på lag och till grundlagens 2 § 3 mom. och önskar sig preciseringar även i detta avseende, vilket betyder att den formella rättsstatligheten är åtminstone på något sätt med i bilden, förstärkt genom anmärkningar om tjänsteansvarets förverkligande enligt grundlagens 118 §. I detta sammanhang upprepar grundlagsutskottet att grundlagens bestämmelser om förvaltningens lagbundenhetsprincip samt statens och tjänstemännens ansvar ger uttryck för principen om tjänstemannaförvaltning. Grundlagsutskottet uppmanade dessutom social- och hälsovårdsutskottet att noga granska hur de föreslagna kriterierna och algoritmerna i de automatiserade metoder som eventuellt tillämpar dem förhåller sig till lagen om offentlighet i myndigheternas verksamhet, och vid behov precisera regleringen. Detta lagförslag, för vilket grundlagsutskottet förutsatte en mängd ändringar för att vanlig lagstiftningsordning kunde användas, förföll emellertid på grund av regeringen Sipiläs avgång. Av den orsaken är det omöjligt att fastställa hur en slutgiltig lagstiftning om socialtrygghets- och försäkringslagstiftningen skulle ha påverkats av grundlagsutskottets anmärkningar beträffande automatiserade beslut.

Däremot hann riksdagen slutbehandla regeringens proposition RP 298/2018 rd med förslag till patientförsäkringslag och till vissa lagar som har samband med den. Synpunkterna om automatiserat individuellt beslutsfattande som framställs är likartade som i ovannämnda utlåtande, vilket förklaras av att behandlingen av de två propositionerna ägde rum parallellt. Det är därför också möjligt att få fram utfallet av grundlagsutskottets tolkningar kring automatiserat beslutsfattande.

Grundlagsutskottet hänvisar till sitt utlåtande GrUU 62/2018 rd om behandling av personuppgifter i migrationsförvaltningen (se nedan) och upprepar en del synpunkter från det utlåtandet beträffande principerna för god förvaltning enligt grundlagens 21 § och bestämmelserna i grundlagens 118 § om tjänsteansvar. Det första lagförslaget anger emellertid inte detaljerat i vilka situationer eller inom vilka ärendegrupper behandlingen kan ske helt automatiskt, vilket grundlagsutskottet anser att leder till preciseringsbehov beträffande regleringen: ”Det måste anges exaktare på vilka grunder ärenden kan avgöras genom automatiserat beslutsfattande.” Här återknyts tematiken till materiell rättsstatlighet.

De vanliga konstaterandena om att avsikten med förslaget inte är att avvika från förvaltningslagens bestämmelser om till exempel hörande av part och andra i grundlagens 21 § 2 mom. tryggade principer om god förvaltning upprepas, men utskottet anser att förslaget bör preciseras i detta avseende på ett sätt som återknyter till den processuella rättsstatligheten. Grundlagens 21 § och kravet i grundlagens 2 § 3 mom. om att offentlig maktutövning ska grunda sig på lag återknyts igen till observationen att kraven på god förvaltning eller parternas rättstrygghet inte får äventyras ens i masshantering, vilka dimensioner bör enligt utskottet preciseras för att det första lagförslaget ska kunna behandlas i vanlig lagstiftningsordning. Därtill knyts förvaltningens lagbundenhet enligt grundlagens 2 § 3 mom. vid automatiserat beslutsfattande igen till tjänsteansvarets förverkligande, och utskottet konstaterar att grundlagens bestämmelser om förvaltningens lagbundenhetsprincip samt statens och tjänstemännens ansvar ger uttryck för principen om tjänstemannaförvaltning. Detta föranleder preciseringskrav i förhållande till det beredande fackutskottet, och bland annat förutsätts formuleringar om att ärendena behandlas i enlighet med de allmänna förvaltningslagarna och att de som behandlar ärendena handlar under tjänsteansvar.

Avslutningsvis upprepar grundlagsutskottet sina tidigare åsikter om att lagen bör innehålla bestämmelser om offentliggörandet av programkoden och dess offentlighet och att det beredande fackutskottet bör utreda hur de föreslagna kriterierna och algoritmerna i de automatiserade metoder som eventuellt tillämpar dem förhåller sig till lagen om offentlighet i myndigheternas verksamhet med tanke på eventuella preciseringar i regleringen. Samtidigt upprepar grundlagsutskottet sina uppmaningar till statsrådet att låta göra ”en utredning för att undersöka hur regleringen av automatiserat förvaltningsförfarande och beslutsfattande uppfyller de krav som följer av förvaltningens lagbundenhet, offentlighetsprincipen och rättsprinciperna för förvaltningen, som ligger till grund för en god förvaltning, samt hur rättstryggheten tillgodoses och tjänstemännens ansvar förverkligas”. Med andra ord, grundlagsutskottet efterlyser en principiell och helhetsmässig syn på regleringen av automatiserat beslutsfattande som förhindrar att det uppstår varierande och situationsbunden reglering av automatiserat beslutsfattande som får olika utfall beroende på regleringsområdet.

Enligt utredning som social- och hälsovårdsutskottet noterade i sitt utlåtande ShUB 38/2018 rd ”fattar Patientförsäkringscentralen för närvarande inga automatiserade beslut. Syftet med bestämmelsen är att ge centralen möjlighet att fatta automatiserade beslut i framtiden. De automatiserade besluten kommer att gälla skadelidandes ersättningsanspråk som är positiva och direkt bygger på en faktura eller ett meddelande som visats upp.” Med hänvisning till grundlagsutskottets utlåtande konstaterar social- och hälsovårdsutskottet att det vid tidpunkten för behandlingen av lagförslaget inte var ”ändamålsenligt att ta in en bestämmelse om automatiserade beslut i patientförsäkringslagen”, när behandlingen av proposition RP 52/2018 rd om ändring av socialtrygghets- och försäkringslagstiftningen med anledning av EU:s allmänna dataskyddsförordning fortfarande pågick i riksdagen (bara för att förfalla p.g.a. regeringens avgång), men det inverkade inte på behandlingen av propositionen om patientförsäkringslagen (948/2019), som godkändes utan bestämmelse om automatiserat beslutsfattande. Social- och hälsovårdsutskottet ansåg att noggrann beredning krävs för att precisera bestämmelsen, en bestämmelse som man alltså ansåg att inte omedelbart behövs, varpå social- och hälsovårdsutskottet föreslog att bestämmelsen inte tas in i den föreslagna lagen. Grundlagsutskottets synpunkter föranledde social- och hälsovårdsutskottet således att föreslå att den i propositionen förekommande bestämmelsen om automatiserat beslutsfattande ska utgå.

3.3 Skarpa ställningstaganden om automatiserade beslut inom migrationsförvaltningen

Automatiserat individuellt beslutsfattande i förvaltningsärenden prövades ingående i grundlagsutskottets utlåtande GrUU 62/2018 rd till förvaltningsutskottet med anledning av regeringens proposition RP 224/2018 rd till riksdagen med förslag till lag om behandling av personuppgifter i migrationsförvaltningen och till vissa lagar som har samband med den. Lagförslaget innehöll en bestämmelse om att beslut som Migrationsverket fattar i ärendehanteringssystemet för utlänningsärenden kan fattas genom ett förfarande som endast grundar sig på automatiserad behandling, om inte annat föranleds av ärendets art eller omfattning, kravet på lika bemötande, barnets bästa eller andra särskilda skäl. Den registrerade skulle ha rätt att få en redogörelse för ett individuellt beslut när beslutet i hans eller hennes sak har fattats genom helautomatisk behandling, dvs. utan att en fysisk person behandlar ärendet, men lagförslaget angav inte detaljerat i vilka situationer eller inom vilka ärendegrupper behandlingen kan ske helt automatiskt.

Grundlagsutskottet hänvisade till lagbundenhetsprincipen i grundlagens 2 § 3 mom., principerna om god förvaltning, särskilt rätten att bli hörd i grundlagens 21 § 2 mom. och till grundlagens 118 § om tjänsteansvar och önskade preciseringar särskilt med avseende på rätten att bli hörd. ”Utskottet har med hänsyn till 21 § och kravet på att offentlig maktutövning ska grunda sig på lag noterat att man inte ens i en masshantering får äventyra kraven på god förvaltning eller parternas rättstrygghet (…). Förvaltningsutskottet bör därför fästa särskilt avseende även vid rättstrygghetsaspekten.” Av denna orsak ansåg grundlagsutskottet att preciseringar av bestämmelserna i 20 § i lagförslaget är en förutsättning för att det första lagförslaget ska kunna behandlas i vanlig lagstiftningsordning. Utskottet hänvisade till 2 § 3 mom. i grundlagen, enligt vilken all utövning av offentlig makt ska bygga på lag och i all offentlig verksamhet ska lag noggrant iakttas och ansåg, i kombination med grundlagens 118 § om tjänsteansvar och i linje med tidigare praxis, att grundlagens bestämmelser om förvaltningens lagbundenhetsprincip och om statens och tjänstemännens ansvar ger uttryck för principen om tjänstemannaförvaltning. Därtill hänvisade utskottet till sin tolkningspraxis, där det ansett ”att det för att kraven på rättssäkerhet och god förvaltning ska anses vara uppfyllda bland annat förutsätts att ärendena behandlas i enlighet med de allmänna förvaltningslagarna och att de som behandlar ärendena handlar under tjänsteansvar”. Grundlagsutskottet förutsatte att bestämmelserna om automatiserade enskilda beslut preciseras för att lagförslaget skulle kunna behandlas i vanlig lagstiftningsordning.

Efter att ha konstaterat att ett indirekt tjänstemannaansvar för de enskilda besluten är otillräckligt om ansvaret kanaliseras genom de personer som kommer att bestämma reglerna för handläggningen vid automatiserat beslutsfattande och som har den faktiska rätt och kompetens att ändra en regel som lett till ett visst beslut, fick förvaltningsutskottet uppmaningen att granska tjänsteansvarets inriktning och vid behov precisera bestämmelsen. Förvaltningsutskottet fick också uppmaningen att fästa uppmärksamhet vid programkodens och algoritmens offentlighet i linje med grundlagsutskottets tidigare praxis.

Ett centralt budskap i utlåtandet är emellertid att ”automatiserat beslutsfattande förefaller inkludera ett flertal frågor som inte har reglerats i de allmänna lagarna för förvaltningen”. Denna observation leder till en uppmaning till regeringen att utreda detta och till behov av en översyn av den allmänna lagstiftningen på området. ”I utredningen bör man undersöka hur regleringen av automatiserat förvaltningsförfarande och beslutsfattande uppfyller de krav som följer av förvaltningens lagbundenhet, offentlighetsprincipen och rättsprinciperna för förvaltningen, som ligger till grund för en god förvaltning, samt hur rättstryggheten tillgodoses och tjänstemännens ansvar förverkligas.”

Det verkar som om grundlagsutskottet i detta sammanhang tänkt framför allt på processuell och materiell rättsstatlighet. På grund av att regeringen Sipilä avgick förföll behandlingen av lagförslaget, vilket ledde till att den efterföljande regeringen Rinne/Marin avgav en ny proposition till migrationslagstiftning, regeringens proposition RP 18/2019 rd med förslag till lag om behandling av personuppgifter i migrationsförvaltningen och till vissa lagar som har samband med den. Grundlagsutskottets utlåtande GrUU 7/2019 rd med anledning av denna nya proposition tog på nytt ställning till automatiserat individuellt beslutsfattande. Utskottet konstaterade att det inte i sig har något att anmärka på de mål som eftersträvas med automatisering av beslutsfattandet, men bedömer det i skenet av principerna för god förvaltning enligt grundlagens 21 § och bestämmelserna i grundlagens 118 § om tjänsteansvar. Bl.a. upprepar utskottet sitt påpekande att masshantering, med hänsyn till grundlagens 21 § och kravet på att offentlig maktutövning ska grunda sig på lag, inte får äventyra kraven på god förvaltning eller parternas rättssäkerhet. Av den orsaken anser utskottet att automatiserat beslutsfattande inte lämpar sig för sådant administrativt beslutsfattande som förutsätter att beslutsfattaren utövar omfattande prövningsrätt.

Grundlagsutskottet kritiserar den föreslagna regleringen för vaghet. ”I lagförslaget fastställs inte vilka ärenden som ska behandlas uttryckligen i ärendehanteringssystemet för utlänningsärenden. Det förblir därför med hänsyn till i propositionen föreslagna 21 § oklart vilka ärenden det automatiska beslutsfattandet över huvud taget kan gälla.” De exempel som nämns i motiveringen till propositionen innehåller vissa ärendegrupper (t.ex. beviljande av medborgarskap som avgörs på ansökan), men det skulle i slutändan vara omständigheterna i det enskilda fallet som avgör om det går att ta fram ett beslut i ärendet helt på automatisk väg, dock inte ärenden där det förutsätts helhetsbedömning gällande till exempel tillräcklig försörjning eller den risk som sökanden orsakar för allmän ordning eller säkerhet. Sådana krävande helhetsbedömningar skulle alltid avgöras av en fysisk person. Avgränsningarna som skulle gälla föreslagna 21 § anges dock inte i förslaget till bestämmelsen, utan enbart i motiveringarna till bestämmelsen. Grundlagsutskottet kan i detta sammanhang anses resonera i termer av materiell rättsstatlighet, då det förutsätter att ärendegrupper inom vilka automatiserat beslutsfattande kan användas ska framgå av substanslagstiftning.

Grundlagsutskottet kommer också in på området för processuell rättsstatlighet när det anser ”att automatiserat beslutsfattande i det regleringssammanhang som nu granskas är möjligt endast när ett yrkande som inte gäller en annan part godkänns”, dvs. när positiva beslut fattas. En uppenbar onödighet att höra parten kan således inte läggas in som ett kriterium som möjliggör användning av automatiserat beslutsfattande, utan användningen av automatiserat beslutsfattande bör begränsas till beslutsfattande där ett yrkande som inte rör en annan part godkänns. Utskottet formulerar också ett alternativ som går ut på en väsentlig precisering av 21 § i lagförslaget så att bestämmelsen anger de situationer som hör till området för den senare satsen i 34 § 2 mom., 5 punkten, i förvaltningslagen och där beslutsfattandet inte förutsätter att myndigheterna utövar prövningsrätt. ”En sådan precisering eller avgränsning av bestämmelserna om automatiska individuella beslut är en förutsättning för att lagförslag 1 ska kunna behandlas i vanlig lagstiftningsordning.” Även processuell rättsstatlighet var således ett centralt tema vid behandlingen av lagförslaget.

Den väsensskillnad mellan mänskligt beslutsfattande genom tjänsteinnehavare och automatiserat beslutsfattande som framkommer t.ex. genom ovan anförda materiella rättsstatlighetstankar understryks av att grundlagsutskottet återkommer till tjänsteansvarets konkretisering vid automatiserat beslutsfattande. Enligt förslagets 21 § 3 mom. skulle Migrationsverkets överdirektör svara för förfarandet vid automatiserat beslutsfattande och för automatiserade individuella beslut och förslaget var till denna del uppenbarligen ett försök att rätta sig efter den kritik om tjänsteansvarets fördelning som utskottet framfört i det tidigare utlåtandet. I detta sammanhang hänvisar utskottet till lagbundenhetsprincipen i grundlagens 2 § 3 mom. enligt vilken all utövning av offentlig makt ska bygga på lag och att lag ska noggrant iakttas i all offentlig verksamhet. Denna allmänna hänvisning inom området för formell rättsstatlighet kombineras i utlåtandet med grundlagens 118 § om ansvar för ämbetsåtgärder, som inbegriper både straffrättsligt och skadeståndsrättsligt ansvar, men också med den straffrättsliga legalitetsprincipen i grundlagens 8 § med särskilt krav på lagens exakthet, dvs. att brottsrekvisitet ska anges tillräckligt exakt så att det utifrån en bestämmelses ordalydelse går att förutse om en viss åtgärd eller försummelse är straffbar. ”Grundlagsutskottet anser det vara klart att överföringen av beslutsfattandet till automatisk behandling inte får leda till att grundlagens bestämmelser om tjänsteansvar förlorar sin betydelse. I det avseendet anser grundlagsutskottet att propositionen är problematisk.” Enligt utskottet lägger den föreslagna bestämmelsen ansvaret ”enbart på överdirektören och stärker i strid med 118 § i grundlagen uppfattningen att andra personer än överdirektören inte kan ha tjänsteansvar”. Utskottet frågar sig emellertid också ”vilka ämbetsåtgärder Migrationsverkets överdirektör i själva verket ansvarar för och vilka ämbetsåtgärder han eller hon i praktiken kan ställas till svars för”, när besluten fattas automatiserat i datasystemet utan direkt och omedelbar mänsklig kontroll över besluten. Riktandet av tjänsteansvaret till överdirektören, som ansvarar för den allmänna ledningen av ämbetsverket, ”ter sig enligt grundlagsutskottet som ett imaginärt och konstlat arrangemang”. I stället anser utskottet att det automatiserade beslutsförfarandet måste, av skäl som beror på grundlagens 118 §, vara noggrant övervakat och juridiskt kontrollerbart, inklusive kopplingen till tjänstemännens ansvar för ämbetsåtgärder, för att det första lagförslaget ska kunna behandlas i vanlig lagstiftningsordning. På vilket sätt tjänsteansvaret i anslutning till automatiserat beslutsfattande ska ordnas i ljuset av detta utlåtande och det föregående utlåtandet GrUU 62/2018 förblir emellertid oklart.

Grundlagsutskottet återkommer också till sina tidigare uppmaningar om att allmän lagstiftning om användning av automatiserat beslutsfattande borde beredas och stiftas. I beredningsarbetet ”bör man undersöka hur regleringen av automatiserat förvaltningsförfarande och beslutsfattande uppfyller de krav som följer av förvaltningens lagbundenhet, offentlighetsprincipen och rättsprinciperna för förvaltningen, som ligger till grund för en god förvaltning, samt hur rättstryggheten tillgodoses och tjänstemännens ansvar förverkligas.” Här framskymtar således både den formella och den processuella rättsstatligheten. Samtidigt som behovet av allmän lagstiftning framhålls innan särbestämmelser inkluderas i den materiella lagstiftningen så konstaterar utskottet att bestämmelserna i dataskyddsförordningen som utarbetats med tanke på skyddet för personuppgifter ”inte utgör en tillräcklig grund för automatiserat beslutsfattande med avseende på principerna för god förvaltning och rättsskyddet inom förvaltningen”.

Det som grundlagsutskottet föreslog mot bakgrunden av sin konstitutionella bedömning var att den tydligaste lösningen är att den föreslagna bestämmelsen i 21 § stryks ur lagförslaget. Om förvaltningsutskottet emellertid vid utskottsbehandlingen skulle gå in för att ändra bestämmelserna om automatiserade individuella beslut så att de överensstämmer med grundlagen, ansåg grundlagsutskottet att utkastet till betänkande om propositionen på nytt ska föreläggas grundlagsutskottet för behandling. Det här framfördes som en förutsättning för att det första lagförslaget ska kunna behandlas i vanlig lagstiftningsordning. Förvaltningsutskottet förordade i sitt betänkande FvUB 10/2020 rd, bl.a. med hänvisning till grundlagsutskottets utlåtande, att förslaget till 21 § stryks ur lagförslaget. Den av riksdagen godkända lagen om behandling av personuppgifter i migrationsförvaltningen (615/2020) stiftades följaktligen utan den konstitutionellt problematiska bestämmelse som ingick i propositionen som förslag till 21 § med bestämmelser om automatiserat beslutsfattande. De skyddsmekanismer som regeringen föreslagit i sina propositioner gällande enskilda lagar var uppenbarligen otillräckliga i ett sådant samhälleligt sammanhang där automatiserat beslutsfattande vid myndigheter var stadd i ökning.

Vid stiftandet av denna lagstiftning uppmanade grundlagsutskottet statsrådet att omsorgsfullt och med iakttagande av ett gott lagberedningsförfarande utreda behoven att ändra den allmänna lagstiftningen om automatiserat beslutsfattande, och vid behov bereder en allmän lagstiftning om automatiserat beslutsfattande som eventuellt kan preciseras genom speciallagar som beaktar respektive förvaltningsområdes särdrag. Här ligger fokus uppenbarligen på både processuell och materiell rättsstatlighet. Utskottet upprepade också sina tidigare iakttagelser om sådana algoritmers offentlighet som används vid myndighetsverksamheten och betonade att ”ett korrekt offentliggörande av algoritmen i en form som är begriplig för enskilda förutsätter att lagen innehåller en exakt och avgränsad definition av vad som avses med automatiserat beslutsfattande genom en algoritm”.

3.4 Sammanfattande synpunkter om grundlagsprövningen

Uppfattningen om krav som grundlagen ställer på automatiserat beslutsfattande vid myndigheter i individuella fall började formas i slutet av 2010-talet och uppvisar en ökande nivå som framhåller olika dimensioner av rättsstatlighet. Den formella rättsstatligheten är ständigt närvarande genom betoningen av att automatiserat beslutsfattande vid myndigheter bör regleras genom lag eftersom sådant beslutsfattande innebär utövning av offentlig makt på det sätt som grundlagens 2 § 3 mom. avser. Denna kategori av rättsstatligheten är emellertid inte särskilt framträdande i grundlagsutskottets tolkningspraxis.

Tyngdpunkten av argument med kopplingar till rättsstatligheten finns inom området för processuell rättsstatlighet, där grundlagens 21 § och principerna om god förvaltning blir centrala, och också inom området för materiell rättsstatlighet, där observationerna om förbättringar i den föreslagna lagstiftningens innehåll gäller de innehållsmässiga grunder som lagbestämmelserna om automatiserat beslutsfattande bör uppvisa.[11] Rättsstatligheten i anslutning till automatiserat beslutsfattande förverkligas därmed framför allt i ett processuellt och materiellt avseende.

Samtidigt är det väldigt tydligt att grundlagsutskottet önskar sig ett mera holistiskt angreppssätt i förhållande till reglering av denna nya teknik för beslutsfattande. Det räcker inte med att skriva in varierande bestämmelser om automatiserat beslutsfattande i substanslagstiftning, utan man får anse att utskottet förutsätter, att allmän förvaltningsrättslig lagstiftning om automatiserat beslutsfattande uppstår, sådan allmän lagstiftning som tryggar god förvaltning, tjänsteansvar och öppenhet vid automatiserat beslutsfattande. Först efter att sådan lagstiftning uppstått och när parametrarna för användning av automatiserat beslutsfattande klarnat kan bestämmelser om ibruktagande av sådana beslutsförfaranden införas i substanslagstiftningen. Som en konsekvens av denna hållning pågår i skrivande stund, hösten 2021, beredning av allmän lagstiftning om automatiserat beslutsfattande och ändringar i offentlighetslagstiftningen vid justitieministeriet och beredning av lagstiftning om tjänstemannaansvarets konkretisering i anslutning till automatiserat beslutsfattande vid finansministeriet.

[11] Det här verkar betyda att man nästan i varje förekommande fall av ibruktagande av automatiserat beslutsfattande borde också förankra det i den relevanta materiella lagstiftningen och därmed anpassa det automatiserade beslutsfattandet till den materiella beslutsmiljö där besluten fattas.

Grundlagsutskottets stränga tolkningar om de förutsättningar som gäller för lagbestämmelser om automatiserat beslutsfattande har i praktiken stoppat sådana separata lagförslag som skulle innebära ändringar i substanslagstiftningen på ett sådant sätt att möjligheten till automatiserat beslutsfattande införs. Utvecklingen av lagstiftningen kring automatiserat beslutsfattande väntar således på den allmänna lagstiftning, tjänstemannalagstiftning och offentlighetslagstiftning som grundlagsutskottet förutsatt att uppstår innan man kan börja förse substanslagstiftningen med särskilda bestämmelser om automatiserat beslutsfattande.

Eftersom grundlagsutskottet inte granskar varje lagförslag med avseende på eventuella statsförfattningsrättsliga komplikationer, kan det emellertid hända att riksdagen lyckats stifta substanslagstiftning med bestämmelser om automatiserat beslutsfattande. Det har inträffat i åtminstone ett fall, nämligen avseende regeringens proposition RP 153/2018 rd med förslag till lag om ändring av den temporära lagen om finansiering av hållbart skogsbruk (34/2015), som är i kraft till slutet av 2023. Regeringen föreslog att lagen skulle få en bestämmelse i en ny 31 a § som möjliggör automatiserat beslutsfattande vid beviljande av stöd för tidig vård av plantbestånd, tidig vård av ungskog och vitaliseringsgödsling samt vid beslut om det slutliga stödbeloppet. Jord- och skogsbruksutskottet föreslog en för denna artikels vidkommande oväsentlig precisering av bestämmelsen, men bestämmelsen godkändes i det stora hela i den form som regeringen föreslagit och antar idag följande utformning:

> 31 a § *Automatiserade beslut*
> Med avvikelse från 14 § 3 mom. i lagen om Finlands skogscentral (418/2011) får beslut om beviljande av stöd för tidig vård av plantbestånd, tidig vård av ungskog och vitaliseringsgödsling och beslut om det slutliga stödbeloppet fattas automatiskt. Med avvikelse från vad som i 6 § 1 mom. i denna lag föreskrivs om att det arbete som får stöd ska vara ändamålsenligt både ekonomiskt och med tanke på bevarandet av skogarnas biologiska mångfald, ska dessa omständigheter beaktas vid beviljandet av stöd endast till den del som uppgifter för att bedöma ändamålsenligheten av det arbete som ska stödjas finns tillgängliga elektroniskt vid det automatiserade beslutsfattandet.

Regeringen hade i sin proposition i viss mån bedömt förslagets förhållande till grundlagen och nämnt principen om god förvaltning i förhållande till förvaltningslagens bestämmelser. Bestämmelsen om automatiserat beslutsfattande är visserligen placerad i en riksdagslag, så den formella rättsstatlighetens minimikrav torde vara uppfyllda, och därtill förekommer innehållsmässiga aspekter (ekonomisk ändamålsenlighet, ändamålsenlighet beträffande skogarnas biologiska mångfald) hos beslutsfattandet som tillgodoser vissa behov inom ramen för den materiella rättsstatligheten. Det är emellertid uppenbart att den godkända lagbestämmelsen saknar alla de dimensioner av processuell rättsstatlighet som grundlagsutskottet ansett vara centrala för tryggandet av grundlagsenligheten hos den lagstiftning som gäller automatiserat beslutsfattande. Dessutom syns det på basis av bestämmelsens andra sats att systemet för det automatiserade beslutsfattandet i detta fall kan behöva utföra operationer som innefattar prövning i och med att det i begränsad utsträckning kan bli fråga om att bedöma ändamålsenlighetsfrågor.

Det är kanske ett misstag att en lagbestämmelse av detta slag kunde bli godkänd, men eftersom utskottsbetänkandet är daterat i mitten av december 2018, är det sannolikt att medvetenheten om de konstitutionella problem som bara några månader senare kopplades till automatiserat beslutsfattande inte ännu var allmänt kända och på det sättet akuta, att lagberedarna och riksdagens utskottspersonal skulle ha börjat begära in utlåtande från grundlagsutskottet. Det är också möjligt att den aktuella lagens temporära natur spelade in som en omständighet som minskade behovet av prövning i grundlagsutskottet av den aktuella bestämmelsen om automatiserat beslutsfattande.

4 Avslutning: formell, processuell och materiell rättsstatlighet

De tolkningar som de högsta laglighetsövervakarna och framför allt riksdagens grundlagsutskott framfört angående automatiserat beslutsfattande har mer eller mindre stoppat all sådan särskild lagstiftningsverksamhet som kunde syfta till att skapa automatiserat beslutsfattande genom bestämmelser i substanslagstiftningen. Däremot kan Folkpensionsanstalten och Skatteförvaltningen fortsätta att använda de automatiserade system för förvaltningsbeslut som de haft sedan tidigare i väntan på allmän lagstiftning, även om dessa myndigheter säkert redan modifierat vissa problematiska detaljer i sin ärendehantering.

För närvarande pågår lagberedning med avseende på allmän lagstiftning om användning av automatiserat beslutsfattande inom förvaltningen,[12] tjänsteansvarets konkretisering vid användning av automatiserat beslutsfattande och algoritmens och programkodens offentlighet och transparens.[13] Det är sannolikt att Finland inom loppet av ett par år har ett allmänt regelverk i kraft som definierar det sätt på vilket lagstiftningen tillåter automatiserat beslutsfattande. Sannolikt är därtill att substanslagstiftningen kommer att innehålla särskilda bestämmelser om automatiserat beslutsfattande och dess särdrag inom vissa förvaltningsområden. Formell, processuell och materiell rättsstatlighet kommer att trygga det att lagbundenhetsprincipen observeras genomgående och på hög nivå när automatiserat beslutsfattande används i myndighetsverksamhet. Rättsstatligheten bör nämligen för det automatiska beslutsfattandets del observeras på samma nivå som beträffande tjänstemannaförvaltning. Den formella rättsstatligheten förutsätter att grunden för automatiserat beslutsfattande etableras i lag och den processuella rättsstatligheten förutsätter att det automatiserade beslutsfattandet underlyder samma allmänna principer om god förvaltning som mänskligt beslutsfattande, medan den materiella rättsstatligheten betonar betydelsen av att det automatiserade beslutsfattandets innehåll vid behov anpassas till det ärende som hanteras.

Den finska grundlagen innehåller annorlunda regler än den svenska statsförfattningen om de konstitutionella krav som gäller för automatiserat beslutsfattande, vilket betyder att regelverken i de två länderna kan se olika ut. Den finska grundlagen förutsätter, *de lege ferenda* och i ljuset av ovanstående tolkningar, ganska mycket av regelverket kring automa-

[12] Se Arbetsgruppen för beredning av allmän lagstiftning om automatiserat beslutsfattande inom förvaltningen: *Automatiserat beslutsfattande inom förvaltningen – Utkast till bestämmelser om användningsområde och öppenhet*, 31.5.2021. Helsingfors: Justitieministeriet, 2021.

[13] Ett intressant steg i riktning mot transparens vid användning av artificiell intelligens inom den offentliga förvaltningen utgörs av den uppräkning som Helsingfors stad anslagit på sina nätsidor av tjänster där AI används i kontakten med kommunmedlemmar. Se https://ai.hel.fi/sv/ai-registret/ (besökt 20.10.2021). De fem AI-system som är i användning (chatbottar, biblioteks- och parkeringssystem) verkar inte inbegripa automatiserat beslutsfattande, dvs. dessa AI-system producerar inte förvaltningsbeslut. I det fall att Helsingfors stad i något skede, och när lagstiftningen tillåter det, tar i bruk automatiserade beslutssystem bör man kunna förvänta sig att även sådana anmäls på nätsidan så att kommuninvånarna har möjlighet att på förhand känna till att automatiserat beslutsfattande förekommer.

tiserat beslutsfattande, medan den svenska statsförfattningen kan tänkas ställa lindrigare processuella och materiella krav, även om den formella lagbundenheten syns likartad.

Inger Österdahl

Laws on LAWS (Lethal Autonomous Weapon Systems): The Work of the United Nations and the Swedish Position

1 Introduction

This contribution deals with Artificial Intelligence (AI) in the context of the law in war, or international humanitarian law (IHL). The development of AI has made possible a high degree of automation of weapon systems. Today, some weapon systems are even called 'autonomous'. Many fear that legally irresponsible and unaccountable machines – or robots – will henceforth make and execute arbitrary decisions over the life and death of human beings.[1] Through international legal regulation, this development could perhaps be forestalled, they hope. We will see how the issue of the normative regulation of so-called lethal autonomous weapon systems (LAWS) is approached at the global level in the United Nations (UN) as well as at the national Swedish level.

It is not easy for the states of the world to reach agreement on what specific normative regulation, if any, should apply to the emerging and important phenomenon of LAWS. Moreover, weapons development is usually something that the states, who are able to produce them, often want to be left as little regulated as possible. In this contribution, we will see what has been achieved so far and what may be achieved in the future.

[1] See, for instance, the Stop Killer Robots campaign, <https://www.stopkillerrobots.org> accessed 28 September 2021.

Some legal questions that arise in the global debate on the regulation of LAWS are whether the (old) international humanitarian law is at all applicable to the (new) phenomenon of autonomous weapon systems. Also, if so, how the existing international law should be applied and whether the existing body of law is sufficient or needs to be complemented by additional rules specific to the field of LAWS. If the existing law in war needs to be complemented, the issue arises as to what the content and form should be of the potential new international rules. And by the way, what exactly are 'LAWS', and what is an 'autonomous' weapon, really?

These are big, fundamental and difficult questions for the international community to solve. The current fierce geopolitical struggles do not make the effort less difficult, and simultaneously the tense international relations make the need for an agreement more urgent. This contribution introduces the unfolding international normative work on providing a solution.

There are four fundamental principles of international humanitarian law that will be referred to several times in this article because of their importance for the debate on LAWS. These are the principles of distinction, proportionality, precaution and the principle of not causing superfluous injury or unnecessary suffering. For the sake of clarity, these principles will be presented here very briefly. According to the principle of distinction, the parties to an armed conflict shall distinguish between combatants and the civilian population and shall direct their operations only against military objectives.[2] According to the principle of proportionality, any incidental loss of civilian life or damage to civilian objects shall not be excessive in relation to the concrete and direct military advantage anticipated by the attack.[3] The principle of precaution stipulates that those who plan or decide upon an attack shall take all feasible precautions, with a view to avoiding incidental loss of civilian life and damage to civilian objects.[4] In the relationship between combatants, finally, according to

[2] See Protocol Additional to the Geneva Conventions of 12 August 1949 and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, Article 48. The author wishes to thank the Torsten Söderberg Foundation for making the writing of this contribution possible through grant number RT2/19 *Krigets juridik i den svenska regeringsformen: ett rörligt mål* (The law of war in the Swedish Instrument of Government: a moving target). All translations to English are made by the author.

[3] See Additional Protocol I, ibid., Article 51 (5) (b); see also Article 57 (2) (a) (iii), and Article 57 (1).

[4] See Additional Protocol I, ibid., Article 57 (2) (a) (ii); see also Article 57 (1).

the fourth principle, it is prohibited to employ methods of or means of warfare of a nature to cause superfluous injury or unnecessary suffering.[5]

2 LAWS and the Group of Governmental Experts (GGE)

Lethal autonomous weapon systems (LAWS) are intensely discussed, but there is no generally accepted definition of what they are. The central point is that they can function without human intervention. However, the question is where the line should be drawn between a highly automated system – which is not subject to discussion within the framework of LAWS – and an autonomous system – which is the subject of discussion within LAWS. With respect to a technical system, the terms automatic and autonomous mean the same thing, that is, the system works without human influence. A technical system can have many automated functions, and a complex system with many automated functions is often labelled autonomous.[6]

Depending on what definition of LAWS one uses – a definition that sets the degree of automation so high that no such weapon system yet exists or a definition that stipulates a lower degree of automation for a weapon system to be labelled 'autonomous' – LAWS can be claimed not to exist today or to exist already. It is uncontroversial to claim that there are weapon systems today with highly automated – or autonomous – functions, without these weapon systems necessarily being subject to discussion within the framework of LAWS. Weapons with certain autonomous functions have existed for more than 100 years. The most common highly automated – or autonomous – weapon systems today are different kinds of targeting robots, which were first put into use during and after the Second World War. Different forms of air defence systems with highly automated – or autonomous – functions also exist, for instance.[7] Probably, the discussion on LAWS relates to weapon systems that do not

[5] See Additional Protocol I, ibid., Article 35 (2).

[6] Cf. *Slutrapport: Arbetsgruppen om autonoma vapensystem* (Final report: The working group on autonomous weapon systems), November 2016, Ministry for Foreign Affairs, on file with author.

[7] Cf. *Dödliga autonoma vapensystem: Rapport till Folkrätts- och nedrustningsdelegationen* (Deadly Autonomous Weapon Systems: Report to the International Law and Disarmament Delegation, Ministry for Foreign Affairs), 25 May 2020, on file with author.

yet exist, but that might soon be brought into existence due to the rapid technological development in the field of AI.

All weapons can be used incorrectly without the weapon itself necessarily being regarded as illegal. Concerning LAWS, a key issue has been whether autonomous weapon systems are able to take into account and apply the fundamental rules of international humanitarian law (IHL), i.e. what is often referred to as the laws in war or the *jus in bello*. If LAWS are inherently unable to take into account and apply the fundamental rules of IHL, then the presumption would be that LAWS in themselves are illegal. Inversely, if LAWS are able to take into account and apply the fundamental rules of IHL, then LAWS would not be inherently unlawful.

Since 2014, the issue of LAWS has been dealt with in the United Nations (UN) within the framework of the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (CCW).[8] Before that, the issue had been brought up before the UN Human Rights Council (HRC) by the UN Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions.[9] The UN Special Rapporteur was concerned about the arbitrariness involved in using drones to target non-state actors and how that challenge could be compounded by the use of autonomous technologies.[10] Specifically, the UN Special Rapporteur used the concept of lethal autonomous robotics (LAR), defined as 'weapon systems that, once activated, can select and engage targets without further human intervention'.[11] Moreover, the Special Rapporteur observed that LARs add a new dimension to the distance that modern technology – for societies with access to it, he points out – allows to be put between weapons users and the lethal force they project.[12] In addition to being physically removed from the kinetic action, the UN Special Rapporteur writes that humans would also become more detached from decisions to kill and from the execution of the decisions to kill.[13] In one

[8] Adopted 10 October 1980, 125 States Parties.

[9] Christof Heyns, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, UN Doc. A/HRC/23/47, 9 April 2013.

[10] Ibid., passim.; cf. also Amandeep S. Gill, "The changing role of multilateral forums in regulating armed conflict in the digital age", *International Review of the Red Cross* (2020), p 261–285, 276.

[11] Heyns, supra note 9, Summary.

[12] Ibid., paras. 26–27.

[13] Ibid., para. 27.

of his conclusions, the Special Rapporteur finds that if left too long to its own devices, the matter of life and death will, quite literally, be taken out of human hands.[14] Moreover, coming on the heels of the problematic use and contested justifications for drones and targeted killing, LARs may seriously undermine the ability of the international legal system to preserve a minimum world order, the Special Rapporteur fears.[15] Then the issue of LARs – or, later, LAWS – was moved from the human rights forum of the HRC to the arms control forum of the CCW, where the problematic legal issues identified by the UN Special Rapporteur remain.[16]

In 2013, the States Parties to the CCW decided that the issue of LAWS would be discussed in informal meetings of experts under the rubric of 'questions related to emerging technologies in the area of lethal autonomous weapon systems'.[17] At the fifth review conference of the CCW in 2016, it was decided that a Group of Governmental Experts (GGE) would be set up which would be open to all States Parties to the Convention.[18] About 80 states have participated in the work of the GGE, among which are found the permanent members of the UN Security Council, the EU member states, as well as numerous civil society organisations, academic institutions, the International Committee of the Red Cross (ICRC) and the UN Institute for Disarmament Research (UNIDIR).[19] In 2018 and 2019, the GGE agreed on eleven guiding principles in total in the area of LAWS.[20] Also in 2019, the States Parties to the CCW

[14] Ibid., para. 110.

[15] Ibid.

[16] Gill, supra note 10, p 276.

[17] Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, UN Doc. CCW/MSP/2013/10, 16 December 2013, paras. 32, 18.

[18] Fifth Review Conference of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or Have Indiscriminate Effects, UN Doc. CCW/CONF.V/10, 23 December 2016, p 9, Decision 1.

[19] Cf., for instance, Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, UN Doc. CCW/GGE.1/2021/CRP.1, 8 December 2021, paras. 6–11.

[20] Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, UN Doc. CCW/GGE.1/2018/3, 23 October 2018, para. 21; Group of Governmental Experts of

endorsed the eleven guiding principles, and the GGE was given the mandate to work out recommendations relating to the clarification, consideration and development of aspects of the normative and operational framework of LAWS.[21] The recommendations, had the GGE been able to reach consensus, would have been presented to the States Parties at the Sixth Review conference of the CCW, which took place in December 2021.[22] All decisions within the framework of the CCW, including the GGE, are adopted by consensus.

3 The guiding principles

Eleven guiding principles have been worked out by the GGE.[23] The guiding principles are preceded by a *general introductory declaration*, where the GGE affirms that international law, in particular the UN Charter and IHL, as well as relevant ethical perspectives, should guide the continued work of the GGE.

The *first of the guiding principles* (a) states that IHL continues to apply fully to all weapons systems, including the potential development and use of LAWS.

Thus, the area of LAWS, although relatively new as a particular area of discussion, is not lawless by default, but LAWS are subject to the application of existing international law. The question of the applicability of old law to new weapons has arisen before. When the issue of the legality of the threat or use of nuclear weapons – invented after most of the principles and rules of humanitarian law applicable in armed conflict had already come into existence – came before the International Court of Justice (ICJ) in 1994, by way of a request for an advisory opinion by the UN General Assembly, the Court found that 'there can be no doubt as to

the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, UN Doc. CCW/GGE.1/2019/3, 25 September 2019, para. 16.

[21] Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Final report, UN Doc. CCW/MSP/2019/9, 13 December 2019, para. 31. The guiding principles are contained in Annex III of the Final report of the Meeting of the High Contracting Parties.

[22] Cf. supra note 19, paras. 12, 17.

[23] Cf. ibid.

the applicability of humanitarian law to nuclear weapons'.[24] Not being able to 'conclude with certainty that the use of nuclear weapons would necessarily be at variance with the principles and rules of law applicable in armed conflict in any circumstance', the ICJ nevertheless found that use of such weapons, in fact, seems 'scarcely reconcilable' with respect for the strict requirements of the principles and rules of law applicable in armed conflict.[25] It remains to be seen whether the issue of LAWS will also come before the ICJ.

The *second principle* (b) lays down an important norm from the point of view of the law, namely that human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system, according to the guiding principle. The concept of human control is much discussed in the context of LAWS and in the context, in particular, of the application of IHL. The question is whether IHL could be applied at all in the absence of human control, ultimately, of the activities of the LAWS and thus whether the use of LAWS under those circumstances can at all be lawful. This, in turn, has to do with the way LAWS are defined – i.e. what does autonomous really mean? – which, as we have seen, is a complicated and, so far, unsettled issue.

According to the *third principle* (c) elaborated by the GGE LAWS, human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of LAWS is in compliance with applicable international law, in particular IHL.

The *fourth principle* (d) states that accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control.

[24] UN General Assembly resolution 49/75 K of 15 December 1994; *Legality of the Threat or Use of Nuclear* Weapons, Advisory Opinion, I. C. J. Reports 1996, p. 226, paras. 85–86.

[25] *Legality of the Threat or Use of Nuclear Weapons*, ibid., para. 95, see also para. 97; a Treaty on the Prohibition of Nuclear Weapons was adopted on 7 July 2017, which entered into force on 22 January 2021, 59 states are parties.

Under the *fifth guiding principle* (e), in accordance with States' obligations under international law, in the study, development, acquisition, or adoption of a new weapon, means or method of warfare, determination must be made as to whether its use would, in some or all circumstances, be prohibited by international law. This guiding principle refers to Article 36 of the Additional Protocol I to the Geneva Conventions according to which:

> [i]n the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.[26]

Thus, the text of the fifth guiding principle is basically identical to a provision in a binding international treaty to which an overwhelming majority of the States of the world are party.[27] Consequently, from the point of view of its content, this particular guiding principle could be said to be comparatively legally binding. Sweden refers to this obligation in its commentary on the guiding principles and generally attaches great importance to this provision in the context of the regulation of LAWS, as well as otherwise.[28]

Under the *sixth guiding principle* (f) further, '[w]hen developing or acquiring new weapons systems based on emerging technologies in the area of lethal autonomous weapons systems, physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered'.

In connection with this, according to the *seventh principle* (g), '[r]isk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems'.

Furthermore, in *principle number eight* (h), it is stated that '[c]onsideration should be given to the use of emerging technologies in the area of lethal autonomous weapons systems in upholding compliance with IHL and other applicable international legal obligations'.

[26] Additional Protocol I, see supra note 2.

[27] Currently 174 States are party to the Additional Protocol I.

[28] See further below in sections 4 and 5.

Significantly, under *guiding principle nine*, (i) '[i]n crafting potential policy measures, emerging technologies in the area of lethal autonomous weapons systems should not be anthropomorphized'. We talk about machines, not human beings, when we talk about LAWS.

Perhaps important as a reminder of the most common uses of AI after all, *the tenth principle* (j) states that '[d]iscussions and any potential policy measures taken within the context of the CCW should not hamper progress in or access to peaceful uses of intelligent autonomous technologies'. It is the peaceful uses of AI, not the bellicose ones, which dominate and should dominate the development of the autonomous technologies.

In *principle number eleven* (k) finally, the GGE does its best to maintain the global discussion on the legal norms governing LAWS and to retain the discussion within the UN in particular. According to the eleventh and final guiding principle, '[t]he CCW offers an appropriate framework for dealing with the issue of emerging technologies in the area of lethal autonomous weapons systems within the context of the objectives and purposes of the Convention which seeks to strike a balance between military necessity and humanitarian considerations'. The latter part of this quote has to be understood also as an implicit reference to the Geneva Conventions on the victims of war, which are founded on an effort to balance military necessity and humanitarian concerns.[29] As we saw earlier, direct references to IHL are made in the opening of the guiding principles on LAWS and in the very first guiding principle, among others, marking the particular normative importance of IHL for the area of LAWS.

[29] Geneva Convention I for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 12 August 1949, 196 States Parties; Geneva Convention II for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea, 12 August 1949, 196 States Parties; Geneva Convention III Relative to the Treatment of Prisoners of War, 12 August 1949, 196 States Parties; Geneva Convention IV Relative to the Protection of Civil Persons in Time of War, 12 August 1949, 196 States Parties; Protocol I Additional to the Geneva Conventions (see supra note 2); Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II), 8 June 1977, 169 States Parties.

4 Swedish comments on the guiding principles

All the High Contracting Parties to the CCW were invited by the Chair of the 2020 GGE to submit commentaries on the operationalisation of the GGE LAWS' guiding principles.[30] Sweden is one of the countries that has responded to the invitation by submitting a commentary.[31] Specifically, Sweden comments on nine out of the eleven principles, paying special attention to the first and third principles, which are considered being the most fundamental. After the presentation of Sweden's comments below, the two principles that Sweden does not provide particular comments on (principles six (f) and eight (h)) will also be mentioned.

Sweden considers that the *first principle* (a) on the applicability of IHL to LAWS is a fundamental principle. In order for the principle to be upheld, Sweden states that it is of utmost importance to train and exercise the personnel in national armed forces in international law applicable during armed conflict. Further analysis would be welcome, Sweden considers, regarding the application of existing IHL with respect to possible future autonomous weapons systems.

On the subject of the *second guiding principle* (b) on retained human responsibility, the Swedish commentary reproduces the fundamental contents of IHL and begins by pointing out that the choice of military means and methods for a military operation must be compliant with the relevant rules and regulations on how military means can be used. The Swedish commentary continues by saying that in planning a military operation, a military commander and his/her staff must consider and assess the presence of civilians (principle of distinction), the principle of proportionality, the principle of precautions in attack and the prohibition of causing unnecessary suffering and superfluous injury. 'The use of a weapon that cannot, or will fail to, fulfil these provisions of IHL

[30] *Commonalities in national commentaries on guiding principles*, UN Doc. CCW/GGE.1/2020/WP.1, para. 1, <https://meetings.unoda.org/section/group-of-governmental-experts- gge-on-emerging-technologies-in-the-area-of-lethal-autonomous-weapons-systems-laws-documents-4929-documents-4947/> accessed 19 August 2021.

[31] Swedish Commentary on the Operationalization of the Guiding Principles on LAWS within the CCW, 30 August 2020, Permanent Mission of Sweden, <https://meetings.unoda.org/section/group-of-governmental-experts-gge-on-emerging-technologies-in-the-area-of-lethal-autonomous-weapons-systems-laws-documents-4929-documents-4947/> accessed 19 August 2021.

may not be deployed or used (sic!)', Sweden continues.[32] These provisions make up the fundamental principles of IHL, and any deployment or use of LAWS that would not meet the requirements of IHL would thus be unlawful. Sweden even wishes to ban LAWS that do not meet the requirements of IHL, as we will see below.[33]

The *third guiding principle* (c) states that human-machine interaction should ensure that the potential use of weapons systems based on emerging technologies in the area of LAWS is in compliance with applicable international law, in particular IHL. Sweden considers this to constitute another fundamental guiding principle, in addition to the first one stating that IHL continues to apply fully to all weapons systems. It is also Sweden's position that preserving human control over the use of force is a key objective. Furthermore, military decision-makers and operators need to be in control – both in terms of their understanding of the weapons systems and their ability and skill to control the systems. Also, all weapons systems have to be predictable and reliable so that their human operators always can be certain that the systems will function in accordance with the intentions of the operator. The Swedish commentary continues: In a military context, rules, regulations and procedures should form a hierarchy of instructions for all operations involving weapons. Any complex system must have rigorous handling regulations, including methods for training and procedures for use. Measures to ensure human control should be considered in the entire life cycle of a weapons system. The specific measures will be context dependent. A system's type of target as well as spatial and temporal limits are likely to be important factors, according to the Swedish view.

Moreover, according to Sweden, in the development of regulations, procedures, manuals and training programmes, the human-machine interaction and its limitations need to be taken into account. In the legal weapons review process (under Article 36 of the Protocol I additional to the Geneva Conventions), an analysis must be performed to ensure that it will be possible to use a given weapons system in compliance with IHL. This analysis should include aspects of human-machine interaction and the ways in which they are addressed in manuals and training programmes.

[32] Ibid.

[33] See section 5.

Finally comes perhaps the most relevant Swedish reflection in the commentary concerning the third guiding principle: The more precise requirements of human control in various contexts still need to be analysed, understood in practical terms and agreed upon. The remaining issue here is exactly what Sweden points out, namely that it still remains to analyse, understand practically, and agree upon what 'human control' means. Or, put in other words, one of the crucial elements – 'human control' – of the normative regulation of LAWS on the global level, at least in the view of Sweden, is still entirely undefined.

If one adds to this the fact that the other crucial element of the normative debate on LAWS also remains undefined in the global context, namely the definition of the actual subject of the debate, the definition of 'LAWS' themselves, then the remaining degree of indeterminacy of the normative global discussion on LAWS becomes evident. An agreement among the participating States on what exactly is being discussed is still outstanding. It is important to note, however, this is not an argument against discussing LAWS on the global level; inversely, it should rather be an argument in favour.

With respect to the *fourth guiding principle* (d), saying that accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control, the Swedish commentary refers back to the comments made concerning the first and second guiding principles.

The *fifth guiding principle* (e) states that in accordance with States' obligations under international law, in the study, development, acquisition, or adoption of a new weapon, means or method of warfare, determination must be made as to whether its employment would, in some or all circumstances, be prohibited by international law. Sweden comments that states have an obligation under international law (Article 36 Protocol I additional to the Geneva Conventions) to determine whether the use of a new weapon would be prohibited under international law.[34] We saw a reference earlier by Sweden to that provision in the context of the third guiding principle. With respect to the fifth guiding principle, Sweden states further that in a review in accordance with Article 36, the characteristics of the weapons system are examined, as well as its planned

[34] Cf. supra note 26.

use and other relevant aspects. In case of doubt or scientific uncertainty, the examining entity could request further information and/or apply further test methods, according to the Swedish commentary. The examining entity is then to issue a decision that approves or rejects the weapons system or method under review. It could also issue strict requirements for modifications or limitations that would bring the system in line with the requirements of international law.

Sweden adds that information is available on a number of national legal review systems – the Swedish one among others – that could assist the States Parties to Protocol I additional to the Geneva Conventions, wishing to create a system for legal weapons reviews or to examine an existing system.

The *seventh guiding principle* (g) – the sixth (f) is not commented upon by Sweden – is relatively straightforward and so is the Swedish comment to principle six. Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems, according to guiding principle six. Sweden comments simply that risk assessments are part of the development of all advanced weapons systems. The processes of procurement, maintenance and use of such systems should be controlled by elaborate safety procedures. The procedures should be documented in handbooks on safety from different perspectives, ranging from questions about explosives and ammunition to software quality, according to the Swedish comment.

The *ninth guiding principle* (i) – the eighth (h) is not commented upon – concerns a subject that stimulates the imagination. According to the ninth guiding principle, in crafting potential policy measures, emerging technologies in the area of lethal autonomous weapons systems should not be anthropomorphised. In fact, this is a phenomenon often pointed out in the discussion of LAWS in different fora. Sweden comments that describing technical systems in a non-technical context is a challenging task. Using adjectives normally used to describe human behaviour easily causes confusion and a risk of drawing inaccurate conclusions about technical systems, which do not possess human qualities. To avoid this, only strictly technical terms should be used.[35]

[35] According to the summary drawn up by the Chair of the 2020 GGE – *Commonalities in national commentaries on guiding principles* – several commentaries underscored that weapons can only ever be tools lacking agency and legal personality, that machines are not

In the *tenth guiding principle* (j), stating that discussions and any potential policy measures taken within the context of the CCW should not hamper progress in or access to peaceful uses of intelligent autonomous technologies, a crucial aspect of the discussion of LAWS is addressed. Here, it can be noted that the characterisation 'intelligent' would seem to anthropomorphise the autonomous technologies right from the start, i.e. the designation '"intelligent" autonomous technologies', or 'artificial "intelligence"' (AI) for that matter, would seem to turn the technical phenomena under discussion inherently anthropomorphised. Perhaps the term 'intelligent' as such should be avoided in the context of describing technologies. This, however, is nothing that Sweden brings up in its comment on the operationalisation of the guiding principles.

Sweden comments instead that, although peaceful uses of technology are not within the scope of the CCW, the following may be noted: The overlap between the civilian and military spheres regarding technology development is significant and appears to be increasing. This creates a mutual dependency, according to Sweden. If a new technology is adapted for military use, the requirements for robustness and reliability of the system need to be set very high.

Sweden continues by saying that technological progress in e.g. automation, autonomy, artificial intelligence and digitalisation/computerisation, is normally common to the military and the civilian spheres, although often driven by civilian (commercial) interests. The challenges of ensuring meaningful control are almost the same for technical systems that may be dangerous (civilian applications), and systems designed to be dangerous (weapons), according to the Swedish comment. This complicates, or makes impossible, the prohibition of certain technologies relating to LAWS since the technologies are used in both the civilian and military spheres.

The eleventh and final guiding principle (k) is relatively straightforward and concerns the appropriate framework for the continued international discussions of LAWS. The eleventh guiding principle states that the CCW offers an appropriate framework for dealing with the issue of emerging technologies in the area of LAWS considering the objectives

moral agents, and that policy measures must always address humans, UN Doc. CCW/GGE.1/2020/WP.1, para. 18 <https://meetings.unoda.org/section/group-of-governmental-experts- gge-on-emerging-technologies-in-the-area-of-lethal-autonomous-weapons-systems-laws-documents-4929-documents-4947/> accessed 19 August 2021.

and purposes of the Convention, which seeks to strike a balance between military necessity and humanitarian considerations. Judging from the Swedish comment, Sweden seems to fully agree with this guiding principle. The participation of experts from several relevant disciplines, as well as representatives from states, civil society and industry, provides a richness of perspectives, Sweden says. Looking forward, the work needed to increase the common understanding of the concept of human control in relation to legal, military and technological aspects is a challenge, Sweden continues. Experts from all the States Parties to the CCW need to be part of the effort, including from the Parties who possess the most advanced capabilities in this area, Sweden concludes.

Sweden does not comment on the operationalisation of principles (f) and (h). Furthermore, there are no explanations for the lack of comments on these principles. Perhaps Sweden considers that the comments on the other principles cover the content of principles (f) and (h) as well, or Sweden considers the contents of the latter principles so self-evident and/or easily operationalised that the principles do not need any further comment. As mentioned earlier, principle (f) states that when developing or acquiring new weapons systems based on emerging technologies in the area of LAWS, physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered. Sweden might perhaps consider that this is already included in the process of review of new weapons under Article 36 of TP I. The second guiding principle not commented on by Sweden – principle (h) – states that consideration should be given to the use of emerging technologies in the area of LAWS in upholding compliance with IHL and other applicable international legal obligations. Perhaps Sweden considers that this sounds reasonable enough and does not call for any further comment.

5 Swedish policymaking on LAWS

Three substantial reports have been produced by different working groups at the Swedish Ministry for Foreign Affairs on the subject of LAWS since 2016, when the GGE LAWS was established.[36] The reports present the

[36] *Slutrapport: Arbetsgruppen om autonoma vapensystem* (Final Report: The Working Group on Autonomous Weapon Systems), 2016, cf. supra note 6; *Dödliga autonoma*

phenomenon of LAWS and the problems involved in getting to grips with LAWS from a normative perspective as well as from the point of view of policy-making more generally. The reports directly and indirectly provide the government with support and suggestions for future Swedish policymaking in this area. Some important points in the most recent Foreign Ministry report (from 2021) will be discussed below in relation to the normative regulation of LAWS. The report is entitled "An Effective Ban on Deadly Autonomous Weapon Systems that are Incompatible with the Requirements of International Law".[37]

The purpose of the report is to put together all the standpoints and perspectives of the members of the working group producing the report, who represent different stakeholders in the Swedish discussion of LAWS, and to make concrete proposals on how Sweden could best push the issue of an effective ban on LAWS that are incompatible with the requirements of international law. In addition to persons coming from different government ministries, the stakeholders represented were the civil society in the form of the Swedish Red Cross and the Swedish branch of the Women's International League for Peace and Freedom, the Swedish defence forces and a couple of defence and peace research institutions.

From the normative perspective, the report addresses the way in which human rights and international humanitarian law apply to LAWS generally. The report also deals specifically with the provision in Article 36 of Protocol I additional to the Geneva Conventions concerning the obligation of States Parties to undertake a review of the compatibility with the Protocol or any other rule of international law of any new weapon, means or method of warfare that the States consider acquiring. The report also briefly presents the contents of another report, authored by the Swedish Red Cross, entitled "IHL and gender: Swedish experiences".[38] A proposal that Sweden should pursue the issue of the integration of a gender perspective in the work on LAWS *inter alia* within the framework of CCW

vapensystem: Rapport till Folkrätts- och nedrustningsdelegationen (Deadly Autonomous Weapon Systems: Report to the International Law and Disarmamament Delegation), 2020, cf. supra note 7; and *Ett effektivt förbud mot dödliga autonoma vapensystem som är oförenliga med folkrättens krav* (An Effective Ban on Deadly Autonomous Weapon Systems that are Incompatible with the Requirements of International Law), April 2021, on file with author.

[37] Cf. ibid.

[38] Cecilia Tengroth and Kristina Lindvall (eds.), Stockholm: Swedish Red Cross and Swedish Ministry for Foreign Affairs, 2015.

was included among the proposals for Swedish action on the road to an effective ban on unlawful LAWS as the last item.

In the report, the defence forces themselves emphasise the importance of meaningful human control in the context of LAWS. A necessary condition for the expediency of the defence forces – i.e. the possibility to reach intended effects and only the intended effects – is that decision-makers and system operators have meaningful human control over the military means used to achieve the effects. Not having meaningful human control is thus not militarily justifiable, and a military reality where such control does not exist is not desirable from the perspective of the defence forces. Another consequence of retaining meaningful human control is that decision-makers and operators can be held accountable for achieved effects, positive as well as any undesired or unlawful effects. According to the report further, in the view of the defence forces, any regulation of automated systems with properties that are dangerous to humans should focus on the concept of meaningful human control.

Then, of course, the question arises as to how the concept of meaningful human control could be defined, regulated and operationalised; this question, however, is currently very far from being answered either at the national Swedish level or at the global level. The Swedish Peace Research Institute (SIPRI) and the International Committee of the Red Cross (ICRC) have recently published a study on different possible ways of exercising human control in the context of LAWS.[39] In the study, SIPRI and the ICRC recommend that future discussions on the normative and operational regulation of LAWS should focus on demands for human control. Thus, the views of SIPRI and the ICRC on the regulation of LAWS, to a large extent, would coincide with the views put forward by the Swedish defence forces in the 2021 Foreign Ministry report. As described above, the issue of human control was also raised in the eleven guiding principles (in principle (b) and (c) in particular) and in the Swed-

[39] Vincent Boulanin, Neil Davison, Netta Goussac, Moa Peldán Carlsson, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*, Stockholm: SIPRI, 2020; see also *ICRC Position on Autonomous Weapon Systems*, 12 May 2021, <https://www.icrc.org/en/publication/455001-icrc-position-autonomous-weapon-systems> accessed 21 February 2022; cf. further, for instance, Filippo Santoni de Sio and Jeroen van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account", *Frontiers in Robotics and AI*, vol. 5, article 15, 2018, 1–14, <https://www.frontiersin.org/articles/10.3389/frobt.2018.00015/full> accessed 28 September 2021.

ish Commentary on the Operationalisation of the Guiding Principles on LAWS within the CCW.[40]

On the subject of the work going on in international fora, the 2021 Swedish Foreign Ministry's report on unlawful LAWS points out that the international work ahead is largely dependent on continued work within the framework of the CCW. Another important aspect, according to the Swedish Foreign Ministry report, presumably affecting the future work with the issue of LAWS, is how high LAWS is on the international agenda. The report does not clearly indicate whether the issue of LAWS is high on the international agenda or whether the issue is not so highly placed. It could seem between the lines as if the authors of the Swedish report fear that, in effect, the LAWS issue might not be all that high on the international agenda. Within the disarmament administration of "a number" of governments, the report says, the LAWS issue plays a "rather prominent and important" role.[41] The same goes for those civil society organisations engaged in peace and disarmament, which also affects policy according to the report. In a number of countries, primarily in Europe, governments and parliaments are also engaged in the LAWS issue, according to the Swedish Foreign Ministry report.

When the Foreign Ministry report summarises the Swedish policy on LAWS so far, the first point taken up is that Sweden pushes for an effective ban on LAWS that are not compatible with the requirements of international law.[42] A bit further down the list, it is stated that an effective ban must include as many countries as possible, of course, also those countries trying to develop the weapon, but this does not necessarily mean that these countries must participate actively in the drafting of the ban, according to the Swedish report. The goal of the Swedish government is as broad a consensus as possible. A further point on the same theme is the Swedish view that a broad agreement in the framework of the CCW would increase the possibilities to reach a future effective ban on LAWS that do not fulfil the requirements of international law. The remaining points in the summary by the Foreign Ministry of the Swedish policy correspond quite well with the content of the Swedish Com-

[40] See supra sections 3 and 4, respectively.

[41] "[E]tt antal" and "tämligen framträdande och viktig" respectively, in Swedish.

[42] "Sverige driver på/ska vara ledande för ett effektivt förbud…" in Swedish.

mentary on the Operationalisation of the Guiding Principles on LAWS within the CCW.[43]

Discussing the issue of a ban in further detail later in the report, the working group begins by stating the obvious, namely that the normative work with LAWS so far is still characterised by a lack of clear definitions. In order to reach a ban, it is fundamental that it is clear what is banned, the report states. Four central questions with respect to the achievement of a ban were particularly discussed by the working group: Human control, the form of a ban/regulation, the content of a ban and finally, the best way for Sweden to pursue the issue of a ban.

The Swedish official position is that LAWS that are incompatible with the requirements of international law should be banned. In the annual statement in 2020 of foreign policy by the government in parliament, the Swedish Foreign Minister Ms Ann Linde said that '[w]ithin the framework of the Convention on Certain Conventional Weapons, Sweden is pushing for an effective international ban on lethal autonomous weapons systems that that are incompatible with the requirements of international law'.[44] In 2021, the Swedish government's policy seemed to remain the same, although formulated in a slightly different manner. In 2021, the Foreign Minister, in the statement of foreign policy, said that '[a] future scenario of lethal autonomous weapons systems (LAWS) that do not comply with international law must be avoided. With the objective of an effective international ban, Sweden is actively participating in the important work within the framework of the Convention on Certain Conventional Weapons'.[45]

With respect to *human control*, the Swedish Foreign Ministry's report finds that in a ban on LAWS, a provision on human control will probably be the most important provision. The report observes that the question of human control has been a core issue since the beginning of the debate on LAWS. There is consensus on the question of human responsibility and of a well-functioning human-machine interaction. There are no explicit requirements for human control in IHL. In the view of the Foreign Ministry working group, clear requirements for human control would be an

[43] See supra section 4.

[44] <https://www.government.se/speeches/2020/02/2020-statement-of-foreign-policy/> accessed 20 August 2021.

[45] <https://www.government.se/speeches/2021/02/statement-of-foreign-policy> accessed 20 August 2021.

effective way of setting boundaries for the development and use of LAWS that are not compatible with international law – i.e. unlawful LAWS. The Foreign Ministry working group writes that an increasing number of countries think that the requirement for human control as an element of the eleven guiding principles should be a central theme in a normative and operational framework. Therefore, whether one is considering a regulation or a ban, both could be based on these principles. The Swedish Foreign Ministry report finds that the designation used for human control varies. Sometimes the designation "sufficient" is used, sometimes "meaningful"; in the context of the GGE, the concept "human-machine interaction" (appropriate for the use and capabilities of a particular weapons system) is used instead, according to the Foreign Ministry report.[46] The Swedish Foreign Ministry working group is almost unanimous in the opinion that Sweden should use the concept '"meaningful" human control'. Then the concept of human control would have to be defined in more detail. This work remains to be done at the international level. Superficially, human 'control' over an 'autonomous' weapon system might seem to constitute a contradiction in terms. In reality, it is probably a question of degree; the weapon system will be more or less autonomous and the human control more or less close.

On the subject of the *form* of a prospective prohibition/regulation of LAWS, the global battle lines on the issue of LAWS appear in the report of the Foreign Ministry working group. In effect, these battle lines probably set the boundaries for the development of LAWS in a real sense. Some countries, the report says, support the view that the efforts within the framework of the CCW should be directed towards achieving a legally binding instrument prohibiting LAWS. On the opposite side, there are a number of countries – Russia, the US, India, Japan, Australia, the United Kingdom, Israel, China – who do not see any need for any additional regulation beyond the already existing IHL. Several EU countries have advocated a political declaration and/or a code of conduct, the Swedish Foreign Ministry report observes. According to the report, a declaration and a code of conduct could lead to a new protocol to the CCW that either regulates or prohibits LAWS.[47] A successful negotiation within the framework of the CCW would either lead to a regulation or a ban, which is followed up regularly by the States Parties, and thereby

[46] Cf. the third guiding principle (c) elaborated by GGE LAWS, supra section 3.

[47] There are five protocols already.

the compliance with the regulation or ban could be effectively monitored. Reasonably, it would be the GGE that would be given the task of drafting a proposal for a protocol, the Swedish Foreign Ministry report says. Since the consensus rule applies in the CCW, all States Parties must agree to such a process.

If the work in the CCW is unsuccessful, there might be proposals for the initiation of work, with a view to a convention outside the CCW framework. This work could take two forms, according to the Swedish report. The first option would be a UN convention after a decision by the UN General Assembly. In order to achieve a decision by the UN General Assembly, it is necessary that a number of countries push the issue with priority, the report observes. The second option would be to pursue the work towards a convention outside the UN framework. In order to be successful, such an effort would presuppose a number of strongly committed countries that would also be willing to finance the conferences and carry out the secretarial work. With respect to both of the latter two alternatives, the Swedish Foreign Ministry report points out, it is improbable that any of the militarily and technologically most important countries would get involved.[48] It remains to be seen whether the efforts to achieve a regulation or a ban in any of the fora listed in the report will be successful.[49]

With respect to all the possible avenues for the negotiation of a regulation or prohibition of LAWS, the active participation of the civil society is important, as stated in the Foreign Ministry report.

On the issue of the *content* of a ban, the Swedish Foreign Ministry report states that irrespective of whether the ban would come about in the form of a new protocol to the CCW or in the form of a convention,

[48] The most recent example would be the Treaty on the Prohibition of Nuclear Weapons Nuclear Weapons, cf. supra note 25, adopted within the UN framework; see also the Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction, adopted 18 September 1997, entry into force 1 March 1999, 164 state parties, adopted outside the UN framework.

[49] For the time being the GGE LAWS will continue its work and its efforts to elaborate 'possible measures' in respect of the normative and operational framework of LAWS; the issue was intensely controversial at the Sixth Review Conference of the High Contracting Parties to the CCW (Sixth Review Conference of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, UN Doc. CCW/CONF.VI/11, 10 January 2022, p 9–10, Decision 1).

the design of the ban can build on previous instances of prohibition of certain types of weapons.[50] A protocol to the CCW may possibly be simpler as to form and content in comparison with a convention, the report observes.

The Swedish Foreign Ministry report argues that a ban may contain both positive and negative obligations. Positive obligations stipulate what requirements are placed on the systems and on the use of the systems (that which is prescribed), while negative obligations indicate what is banned (that which is proscribed).

The purpose of the positive obligations is to maintain human control throughout the entire life cycle of the weapons system. Maintaining human control requires a clear chain of order and control by human beings and demands predictability when weapons are used, the report says. The predictability of the mode of operation of a deadly technical system is necessary in order for the system to be compatible with IHL, the Foreign Ministry Report continues. Positive obligations can be drafted so that they are possible to check and follow up; furthermore, positive obligations underline the importance of compliance with IHL, the report states. The negative obligations normally clearly state what is prohibited. The Swedish Foreign Ministry says that it should be possible to state that deadly autonomous weapon systems that cannot respect the principles of distinction, proportionality and precaution, including the prohibition of causing superfluous injury or unnecessary suffering, are banned. However, the report adds that research on LAWS for purposes of defence or protection should not be banned. It is unclear whether this would include research on potentially unlawful LAWS.

On the last of the four central questions with respect to a ban on unlawful LAWS discussed by/within the Swedish Foreign Ministry working group – *the best way for Sweden to pursue the issue of a ban* – the report essentially finds that cooperation with other States is necessary for success, including cooperation at the global level. The Nordic countries, some EU states, Switzerland, and a number of countries in other parts of the world are mentioned in particular. It is pointed out in the report that it is also important to have good contacts with the States that have the greatest capacity to develop LAWS. Since Sweden is an important manufacturer of weapons, it is also pointed out in the report that the States included in the so-called six nations collaboration between the six biggest defence in-

[50] Cf. supra note 48.

dustry nations in the EU (in addition to Sweden: France, Italy, Spain, the United Kingdom and Germany) are also important from the perspective of weapons development.[51]

The working group drafting the Foreign Ministry Report is of the opinion that the best way for Sweden to proceed would be to pursue the issue of a ban on unlawful LAWS within the framework of the CCW and build on what has already been achieved by the GGE.[52] Judging from the report, the prospect of the GGE getting the mandate to draft a new protocol to the CCW, either in the form of a ban or in the form of a '"clear" regulation', is not entirely unrealistic. However, the report ominously points out that there is great uncertainty about when the 2021 review conference of the CCW will take place.

An important point made in the Foreign Ministry report among the proposals for Swedish action on the road to an effective ban on unlawful LAWS is that Sweden should promote the weapons review process under Article 36 of the Protocol I additional to the Geneva Conventions. Since the review of the compatibility with IHL, or any other rule of international law, of the use of any new weapon already constitutes a binding obligation for States under international law, promoting respect for this provision would seem to be a good idea. The report proposes that the weapons review process under Article 36 Additional Protocol I is promoted in the EU as well as in the CCW framework. In the latter case, the weapons review issue should be pursued regardless of the outcome of the discussions on LAWS, the report says, well aware of the difficulties involved in moving forward in the normative work on LAWS. As the Swedish Foreign Ministry report aptly finds, a well-functioning weapons review process under Article 36 on the global scale should also have the capacity to catch LAWS that are incompatible with the requirements of international law, since by definition these are 'prohibited by [Additional Protocol I to the Geneva

[51] Cf. Sveriges överenskommelser med främmande makter (SÖ) 2001:13 Framework agreement between The French Republic, The Federal Republic of Germany, The Italian Republic, The Kingdom of Spain, The Kingdom of Sweden, and The United Kingdom of Great Britain and Northern Ireland concerning measures to facilitate the restructuring and operation of the European defence industry, concluded 27 July 2000, entry into force for Sweden 6 May 2001.

[52] See also Group of Governmental Experts on emerging technologies in the area of Lethal Autonomous Weapons Systems, General Statement by Sweden, Geneva, 3–13 August 2021, <https://meetings.unoda.org/section/firstsession_statements> accessed 28 September 2021.

Conventions] or by any other rule of international law', as stipulated in Article 36. We will see what comes first, a ban on unlawful LAWS or general respect for the weapons review process under Article 36.

6 Conclusion

LAWS (lethal autonomous weapon systems) are here to stay, and the question is whether they need to be regulated internationally and if so, how. This contribution has dealt with the background to the current international normative debate on LAWS, the tentative attempts in the UN to agree on a normative framework for LAWS, and the Swedish position with respect to the emerging normative principles and further policymaking in the area. Widespread disagreement remains at the global level on what would be the appropriate form and content of any new regulation of LAWS. Widespread disagreement also remains on the definition of 'LAWS'. The international normative efforts currently within the framework of the CCW (Convention on Certain Conventional Weapons) are focused on the concept of 'human control'. Retained human control, ultimately, over the otherwise highly automated weapon systems, is a necessary condition for it to be possible to apply existing law – primarily IHL (international humanitarian law) – or any law. Thus, human control has become the hook on which the current discussion hangs.

It might be speculated that the loss of human control over advanced weapon systems would be in no human being's interest, irrespective of how technologically resourceful one's home country is. Still, there is an evident negative correlation between the ability of a country to develop LAWS and its willingness to submit to further international normative regulation of highly automated weapon systems. Conversely, the countries aiming to achieve further international regulation of LAWS are typically those countries lacking the resources to develop LAWS themselves.

Sweden intends to pursue an international ban on LAWS that are incompatible with the requirements of international law. If a ban turns out not to be attainable, another kind of clear international regulation of LAWS might be an alternative.

Sweden also emphasises the importance of the compulsory legal review of new weapons (or means or method of warfare) under Article 36 in Protocol I additional to the Geneva Conventions. A careful national implementation of the legal review under Article 36 would result in a determination as to whether the employment of the new weapon would be prohibited, or not, by IHL or any other rule of international law. In addition to the ban on the use of an unlawful weapon that would potentially follow from such a review under Protocol I additional to the Geneva Conventions, Sweden intends to pursue a ban on the unlawful weapon itself.

Marianne M. Rødvei Aagaard

Digitalisering av undervisningsmaterial och lärarens upphovsrätt[1]

1 Inledning

Kontrasten mellan den tid då lärarens anteckningar försvann genom att sudda tavlan och dagens mångfald i tekniska och pedagogiska verktyg för att skapa undervisningsmaterial är betydande. Utvecklingen har givetvis pågått länge, men under senare år har särskilt tillgången på digitala verktyg gjort att tröskeln för att skapa material som medger en längre tids bevarande blivit låg. Digitaliseringen inom högskolesektorn innebär inte bara att undervisningsmaterial skapas i nya former och på nya sätt, utan också att det skapas i former som tillåter återanvändning och spridning på andra sätt än vad vi tidigare har varit vana vid.[2] De tekniska och pedagogiska frågor som uppkommer i detta sammanhang är givetvis många, men utvecklingen av olika former av digitalt undervisningsmaterial har

[1] Inspiration till ämnet kom under en högskolepedagogisk kurs om digitala kompetenser i undervisningen vid Stockholms universitet hösten 2020, där variationen i uppfattningar om hur lärarens upphovsrätt och lärosätenas förfoganderätt förhåller sig till varandra visade sig vara mycket stor, både bland deltagare och kursansvariga. Stort tack till docent Christina Wainikka och doktorand Annika Blekemo för värdefulla inspel till ett tidigare utkast av texten, och till forskningsamanuens jur.stud. Ludwig Irveland för hjälp med korrektur m.m.

[2] Digitaliserad undervisning är dock inget egentligt nytt, utan sättet att formulera sig har använts länge. Se t.ex. Magnusson Sjöberg, En rättslig ram för e-lärande – integritetsskydd m.m., Liber Amicorum festskrift till Jan Rosén, s. 525 ff. på s. 530 ff. som ger en rad exempel på hur den teknologiska utvecklingen 2016 hade påverkat vår verksamhet, och Morten Rosenmeier, Ophavsretten til undervisningsmateriale. Nogle danske erfaringer, Liber Amicorum festskrift till Jan Rosén s. 695 ff. [Rosenmeier (2016)]. Se också Wolk, Arbetstagares immaterialrätter: rätten till datorprogram, design och uppfinningar m.m. i anställningsförhållanden, 2 uppl., 2008 [Wolk (2008)] s. 212 f. och s. 224 f.

också lett till att frågan om lärarens (skaparens) upphovsrätt aktualiserats på nytt. Att upphovsrätten till det material som skapas å ena sidan tillfaller individen, och att en arbetsgivare å andra sidan gärna anser sig berättigad att förfoga över de verk som skapas inom ramen för en anställning eller ett uppdragsförhållande, är två motstående utgångspunkter som står i ett tydligt spänningsförhållande till varandra.[3] Konflikten har dock hittills bara i begränsad utsträckning ställts på sin spets när det gäller förfoganderätten till undervisningsmaterial skapat av universitetslärare.[4]

Ämnet blir särskilt aktuellt i fall där undervisningsmaterial skapas eller framförs i former som tillåter exemplarframställning och återanvändning utan ändringar eller större insatser från vare sig den skapande läraren eller andra. Återanvändning skulle, om den kunde göras oberoende av lärarens samtycke och utan kompensation till läraren,[5] potentiellt kunna leda till stora besparingar i institutionens ofta hårt pressade ekonomi. Den lärare som har skapat materialet kan emellertid tänkas ha flera invändningar mot ett sådant förfarande. Exempel på invändningar kan vara att tidsåtgången för att skapa och framföra verket varit väsentligt högre än den timavräkning som den egna användningen hittills gett,[6] att läraren själv anser att ändringar eller kompletterande insatser vid sidan av verket be-

[3] Om upphovspersonen är anställd eller en uppdragstagare som utför undervisning på beställning kan tänkas få viss betydelse för frågan om lärosätets förfoganderätt till materialet. Framställningen tar sin utgångspunkt i att analysera den anställda lärarens upphovsrätt, men får till största delen anses ha bäring också för den externa inhyrda föreläsaren, jfr särskilt avsnitt 3.

[4] Se Wolk (2008) s. 220 f. om vilka högskoleanställda som kan stå i en upphovsrättslig särställning. Se också Wolk, Lärarundantaget i omvandling, NIR 2003 s. 415–426. Eftersom innehållet i artikeln i all huvudsak får anses konsumeras av andra upplagan av hennes avhandling, hänvisas i det följande uteslutande till Wolk (2008).

[5] Det sätt som en lärare typiskt sett skulle kompenseras vore genom att den fick tillgodoräkna sig undervisningstimmar för användningen, jfr närmare nedan i avsnitt 3.2.3.

[6] Invändningen är inte egentligen av upphovsrättslig karaktär, men upphovsrätten kan ändock tänkas komma läraren till undsättning vid en sådan invändning. Det kan vara stor variation mellan olika sorters förinspelat material gällande format, innehåll, tidsåtgång för att skapa eller justera för återanvändning etc. Två exempel på var sin sida av en lång skala kan nämnas. Att klicka på ”record” innan en muntlig föreläsning över t.ex. Zoom påbörjas och sedan återanvända inspelningen tar inte mycket mer tid än den tid som gick åt för att hålla föreläsningen. Att strukturera om en föreläsning i kortare delar för att anpassa innehållet till ett nytt medium och lägga tid på inspelning, omtag, redigering och design för att det inspelade materialet ska bli användarvänligt och tåla uppspelning flera gånger, kan däremot ta flera veckor i effektiv arbetstid, givetvis beroende på program, förkunskaper, ämne etc.

hövs för att kvaliteten ska vara tillräckligt hög för att han eller hon ska vilja vidkänna sig det, eller av mer principiell karaktär: Det är jag som har skapat verket, och då är det jag som bestämmer över dess användning!

Digitaliseringen och den teknologiska utvecklingen öppnar för att frågor som tidigare kanske inte var tillräckligt betydelsefulla för att vare sig lärare eller lärosäten på allvar skulle orka bråka om vad som egentligen gällde, plötsligt kan bli av betydelse för båda. Ett exempel kan ges med utgångspunkt i den traditionella föreläsningen. Framförs föreläsningen muntligt med alla åhörare på plats i en sal, ställs den potentiella upphovsrättsliga konflikten mellan lärare och lärosäte sällan på sin spets. Eftersom verket framförs muntligt låter det sig inte lätt återanvändas som sådant vid ett senare tillfälle. Snarare skulle återanvändning då först kräva exemplarframställning i ett format som i sig kräver viss insats, och som läraren kan motsätta sig av andra skäl.[7] Har läraren däremot själv spelat in föreläsningen med ljud och bild, blir frågeställningen om arbetsgivarens återanvändning och förhållandet till upphovspersonens ensamrätt plötsligt mer praktisk.[8] Även frågor om upphovspersonens ideella skydd enligt 1 kap. 3 § upphovsrättslagen kan i dessa fall tänkas uppkomma, men dessa ska inte behandlas närmare här.[9]

[7] Att spela in och återanvända en muntlig föreläsning (med ljud och/eller bild) utgör en exemplarframställning som kräver samtycke, jfr Kielland (2018) s. 52. Inspelning och efterföljande spridning utan samtycke skulle förutom att utgöra ett intrång i upphovspersonens ekonomiska ensamrätt också potentiellt kunna anses som ett intrång i lärarens ideella skydd enligt 1 kap. 3 § upphovsrättslagen, jfr SOU 1956:25 s. 185 om att det även skulle kunna vara problematiskt om en student kunde överlåta egna nedteckningar från en föreläsning till försäljning. Sandgren, Rätten till undervisningen och forskningen, XXVI SULF:s Skriftserie, 2003 [Sandgren (2003)] s. 25–26 utgår ifrån att studenterna fritt kan spela in föreläsningar, men endast använda inspelningarna för enskilt bruk. Se närmare Wolk, Inspelning av föreläsningar – några upphovsrättsliga reflektioner, i SvJT 2009 s. 717 ff. [Wolk (2009)] om andra upphovsrättsliga frågor som kan uppkomma i samband med inspelning av föreläsningar och om exemplarframställning för privat bruk i 2 kap. 12 § upphovsrättslagen. Som Rosenmeier (2016) s. 707 framhåller torde emellertid inspelning förutsätta samtycke på grund av regleringen till skydd för föreläsarens personuppgifter, vilket innebär att läraren både gentemot studenter och arbetsgivare kan förhindra inspelning mot sin vilja.

[8] När läraren föreläser eller på annat sätt framför sitt verk, anses han eller hon som en utövande konstnär enligt 5 kap. 45 § upphovsrättslagen och får därmed också för framförandet en exklusiv rätt att spela in, framställa exemplar och tillgängliggöra framförandet eller en upptagning av det för offentligheten.

[9] Exempelvis kan ny praxis ha kommit till och lagar ändrats efter att inspelningen gjordes, vilket innebär att föreläsningen inte längre är uppdaterad. En återanvändning kan då

I frågan om lärosätenas möjlighet att återanvända material skapat av sina anställda lärare finns tydliga kopplingar mellan arbetsrätten och upphovsrätten.[10] Under 2020 kunde man i olika diskussionsforum för universitetslärare följa engagerade och ibland upprörda diskussioner om tidsåtgång för att skapa förinspelat och annat digitalt tillgängligt material, och hur saker skulle bli *post covid*. Skulle läraren bli överspelad, därför att det då skulle finnas så mycket bra och anpassat digitalt material som lärosätena skulle kunna återanvända utan lärarens samtycke, och utan att ge läraren timmar för det? Skulle läraren få timmar för den tid som skulle komma att behövas för att skapa sådant material, eller bli hänvisad till vanliga nycklar för beräkning av undervisningstid? Hur skulle man förhålla sig till beräkning av undervisningstid om läraren senare valde att återanvända förinspelat material? Frågorna var många, och väsentligt fler än de som nämnts här.

Situationen som uppkom under 2020 innebar också att lärosätenas riktlinjer för hantering av immateriella tillgångar skapade av dess anställda uppmärksammades.[11] Att ha riktlinjer som förtydligar vad som upphovsrättsligt gäller mellan lärosätena som arbetsgivare och lärarna som anställda är givetvis inte i sig dåligt om riktlinjerna är i enlighet med den rättsliga regleringen. Inte sällan är intrycket efter att ha läst

bli kränkande för upphovspersonen, särskilt om inspelningsdatum inte framgår eller har redigerats bort för att underlätta för återanvändningen.

[10] Se närmare nedan i avsnitt 3.2 om anställningsförhållandets särprägel m.m.

[11] Se riktlinjer för Stockholms universitet beslutade 2013 https://www.su.se/medarbetare/organisation-styrning/styrdokument-regelboken/kommunikation-och-samverkan/riktlinjer-avseende-r%C3%A4tten-till-undervisningsmaterial-1.121205, Örebro universitets riktlinjer beslutade 2013 https://www.oru.se/globalassets/inforum-sv/centrala-dokument/styrdokument/personal/anstallning/nyttjanderatten-till-undervisningsmaterial-riktlinjer.pdf, Karlstads universitets riktlinjer beslutade 2017 http://intra.kau.se/dokument/upload/C10B942807b1d25F87rY81BFC7EA/C2017_233policy_nyttiggorande.pdf, Uppsala universitets Riktlinjer avseende intellektuella tillgångar skapade vid Uppsala universitet beslutade 2014, under revision våren 2021, finns inte att tillgå via webben utan inloggning, Lunds universitets riktlinjer beslutade 2017 https://www.medarbetarwebben.lu.se/sites/medarbetarwebben.lu.se/files/allmanna-rad-om-lunds-universitets-nyttjanderatt-till-upphovsrattsligt-skyddat-material.pdf, riktlinjer Linköpings universitet beslutade 2018 https://styrdokument.liu.se/Regelsamling/VisaBeslut/876979, riktlinjer Göteborgs universitet beslutade 2020 (https://pil.gu.se/digitalAssets/1776/1776270_gu-2020-1226-policy-for-universitetets-nyttjanderatt-till-upphovsrattsligt-skyddat-material.pdf, Umeå universitets riktlinjer (utan datum) https://www.aurora.umu.se/utbilda-och-forska/stod-till-utbildning/upphovsratt-och-avtal-om-kopiering-och-delning/anvandning-av-undervisningsmaterial/. Alla länkar besökt senast 2 augusti 2021.

dessa riktlinjer att man inte avsett att uttrycka något nytt, utan endast förtydliga vad som oavsett gäller.[12] En fråga som uppkommer är emellertid om dessa riktlinjer, som i stor utsträckning anger att lärosätet har rätt att nyttja det material som skapats av dess lärare så snart som det har offentliggjorts, verkligen är förenliga med upphovsrättslagen och rättsläget i övrigt.[13]

För att kunna göra en sådan bedömning behöver man först ha en uppfattning om vad som egentligen gäller avseende lärarens upphovsrätt. I detta bidrag ska därför frågan om universitetslärarens upphovsrätt till *undervisningsmaterial* och vilken betydelse denna har för lärosätets möjlighet att använda sig av materialet i verksamheten behandlas.[14]

Eftersom upphovsrätt originärt endast tillkommer *fysiska* personer,[15] är lärosätets eventuella rättigheter till materialet en från upphovspersonen härledd rätt.[16] I framställningen bortses det därför från den (annars) tänkbara möjligheten att upphovsrätten vore gemensam för lärare och lärosäte. I avsnitt 2 nedan behandlas de rättsliga utgångspunkterna för lärarens upphovsrätt och möjliga inskränkningar i denna, medan frågan om lärosätets rätt att nyttja upphovsrättsligt skyddat material skapat av läraren behandlas i avsnitt 3.

[12] Flertalet av de i fotnot 11 nämnda riktlinjerna anger att det uteslutande handlar om universitetets tolkning av vad som oavsett gäller. Trots det finns det viss variation i innehållet mellan de olika riktlinjerna och innehållet är ofta inte helt förenligt med vad som följer av rättsläget, jfr närmare nedan.

[13] En angränsande fråga är om arbetsgivaren, med eller utan fackets medgivande, har möjlighet att utan särskild ersättning eller avtal med den enskilde göra ingrepp i lärarens upphovsrätt. Man kan notera att SACO-S i samband med att Linköpings universitet beslutade om riktlinjer har varit mycket negativa till innehållet och att riktlinjerna också MBL-förhandlats utan att man uppnått enighet. Det vore högst oväntat om facket någonsin skulle ingå kollektivavtal som innebär övergång av upphovsrätt i dessa fall. Se https://www.saco.se/lokala-webbplatser/saco-s-vid-liu/nyheter/liu-vill-nyttja-upphovsrattsskyddat-undervisningsmaterial-en-bakgrund-och-kritik-fran-saco-s-vid-liu/ för en översikt över invändningarna facket framförde (besökt 2 augusti 2021).

[14] Det avgränsas alltså mot en rad angränsande frågor, såsom t.ex. frågor om upphovsrätt vid uppdragsforskning, lärosätets användning av verk i tråd med Bonusavtalet, studenternas rätt att för enskilt bruk framställa exemplar etc.

[15] SOU 1956:25 s. 83, prop. 1960:17 s. 50 f.

[16] Se Karnell, Arbetstagares upphovsrätt, NIR 1969 s. 54–67 [Karnell (1969)], Godenhielm, Arbetstagares upphovsrätt, NIR 1978 s. 321–351 [Godenhielm (1978)] och Kielland, Opphavsrettslige problemstillinger ved digitalisering av undervisning ved universiteter og høyskoler, 2018, [Kielland (2018)] s. 41.

2 Utgångsläge och undantag

2.1 Förutsättningar för upphovsrätt

För att upphovsrätt ska uppkomma krävs att det finns ett *verk* som har skapats av en person genom dennes *personliga insats*. Ett verk måste inte ha en bestämd form, utan kan framföras såväl muntligt som skriftligt, och framställas i analoga eller digitala format.[17] Detta innebär att i princip allt material som en lärare skapar *kan* vara framställningar av verk som skyddas av upphovsrätt.[18] I tillägg kan framförandet *som sådant* skyddas enligt 5 kap. 45 § upphovsrättslagen, vilket ger läraren som utövande konstnär skydd för själva framförandet.

Upphovsrättsligt skydd uppkommer endast för verk som har vad som tidigare kallats verkshöjd, vilket innebär att krav ställs på att alstret ska uppvisa viss originalitet, individualitet och självständighet.[19] Idag är det, mot bakgrund av att EU-domstolen fastställt att verksbegreppet ska tolkas och tillämpas på ett enhetligt sätt, mer aktuellt att tala om ett originalitetskrav.[20] Tröskeln för upphovsrättsligt skydd är dock relativt låg.[21] Det centrala är att verket ska vara uttryck för upphovspersonens egna, intellektuella skapande, och det finns inga krav på vare sig kvalitet eller omfång. Det ställs heller inte krav på visst mått av särprägel och verket i sig behöver inte vara igenkännbart som något skapat av just den aktuella

[17] Se definitionen i 1 kap. 1 § 1 upphovsrättslagen och t.ex. Sandgren (2003) s. 20, Strömholm ”Vem äger forskningen” (2002) [Strömholm (2002)] s. 41, Wolk (2009) s. 719.

[18] Se närmare Wolk (2009) på s. 718 ff. om föreläsningars upphovsrättsliga skydd och historiken bakom nu gällande lagreglering. Se också SOU 2005:95 Nyttiggörande av högskoleuppfinningar s. 239.

[19] Se t.ex. Sandgren (2003) s. 21, Wolk (2009) s. 720, Bernitz, Karnell, Pehrson, Sandgren, Immaterialrätt och otillbörlig konkurrens, 15 uppl., 2020 [Bernitz m.fl. (2020)] s. 50 ff.

[20] EU-domstolen har i avgörandena C-5/08 (Infopaq) och C-683/17 (Cofemel) fastställt att begreppet verk ska tolkas och tillämpas på ett enhetligt sätt. I denna tolkning står originalitetskravet centralt i den meningen att det är upphovsmannens egen intellektuella skapelse och att endast element som ger uttryck för upphovsmannens intellektuella skapande kan kvalificeras som verk. Om det är mest träffande att se det så att verkshöjdsbegreppet tolkats om i och med avgörandena från EU-domstolen, eller om begreppet rentav blivit obsolet och bör bytas ut med ett originalitetskrav finns det skilda uppfattningar om. Att tröskeln för upphovsrättsligt skydd följer av unionsrätten torde dock inte (längre) vara tveksamt.

[21] Se t.ex. Sandgren (2003) s. 21, Wolk (2009) s. 720, Bernitz m.fl. (2020) s. 60.

personen. Inte heller krävs att verket har ett nyhetsvärde, utan också en föreläsning om ett uttjatat ämne kan i princip nå upp till kravet, om den skapats genom lärarens personliga intellektuella insats.[22] Att andra personer, till exempel en annan lärare med motsvarande ämnesinriktning och erfarenhet, med utgångspunkt i samma material *hade kunnat* skapa ett liknande verk, är också utan betydelse. Att det vid ett sådant fall i och för sig kan bli svårt att visa att det är just den personliga insatsen, och inte andras insats, som har lett fram till slutprodukten kan dock inte uteslutas.

Att frågor om lärarens upphovsrätt till undervisningsmaterial inte så ofta aktualiseras, har nog till stor del att göra med vår självbild och vår syn på skapandet av sådant material. Att skriva en bok eller en vetenskaplig artikel uppfattar nog alla som en, åtminstone i viss mån, konstnärlig prestation som leder fram till ett verk. Att utforma en seminarieuppgift, en tentamensfråga eller en föreläsning, väcker inte nödvändigtvis sådana associationer. Vi ser gärna detta som enklare och praktiska uppgifter, som utförs nästan mekaniskt. Seminarieuppgifter och tentamensuppgifter som innehåller beskrivningar av tänkta omständigheter som leder oss till bestämda rättsliga problem förutsätter emellertid fantasi såväl som ämneskompetens och personlig insats, och en tentamenskommentar kan på liknande sätt som en vetenskaplig text kräva noggranna analyser och betydlig pedagogisk insats. Detsamma gäller muntliga föreläsningar, föreläsningsmanus och andra presentationer, oavsett om presentationsbilderna är att anse som enkelt stödmaterial för en muntlig föredragning eller närmare fulltext.[23] Rent principiellt spelar det heller ingen större roll för frågan om verkshöjd om verket ges uttryck i digital eller analog form, men man torde kunna utgå ifrån att den personliga insatsen och tidsåtgången i många fall blir än större om man lägger till den ytterligare dimensionen som det kan innebära att anpassa innehållet för digitala format, och spela in eller på annat sätt digitalisera undervisningsmaterialet.

[22] Elementen i de verk som omfattas av upphovsrätten kan vara sådana att de betraktade separat inte i sig utgör en intellektuell skapelse, men genom valet, dispositionen och kombinationen av dessa element kan upphovspersonen ge uttryck för sin kreativitet på ett originellt sätt och nå ett resultat som utgör en intellektuell skapelse, jfr C-5/08 (Infopaq).

[23] När det handlar om olika format som alla rör en enda föreläsning är det närliggande att det handlar om exemplarframställning av ett och samma verk, men beroende på hur materialet är tänkt att användas av upphovspersonen kan det också handla om olika, självständiga verk. Ur upphovsrättslig synvinkel behöver emellertid ingen exakt gräns dras i detta avseende.

Mot denna bakgrund kan man relativt enkelt tänka sig att en större del av det material som universitetslärare skapar i samband med utförandet av sina undervisningsuppgifter blir skyddat av upphovsrätt, på samma sätt som lärarens läroböcker och vetenskapliga texter blir det. För det material som skapas av läraren i samband med sin undervisning är det normalt utan betydelse vilken sorts material det handlar om, förutsatt att kravet på verkshöjd uppnåtts.[24] Någon skiljelinje mellan föreläsningar och annat som en lärare skapat för sin personliga användning i den egna undervisningen, och vad som skapats på uppdrag från lärosätet eller till användning för flera lärare, behöver då inte dras eftersom den är utan betydelse för frågan om läraren *har upphovsrätt* till materialet. Vilket slags material det är fråga om kan emellertid tänkas ha betydelse för om lärosätet får nyttja materialet utan ett uttryckligt avtal med läraren, jfr närmare nedan i avsnitt 3.

2.2 Upphovsrätten som ensamrätt

Upphovsrätten består av en ekonomisk och en ideell del, och tillkommer enligt 1 § upphovsrättslagen den som har *skapat* ett verk, och rättigheterna är till sin karaktär *ensamrättigheter*.[25] Med detta menas att den som genom sin egen personliga skapande insats exempelvis har författat en text eller planerat, skapat och framfört en föreläsning *ensam* har rätt att nyttja verket på det sätt som omfattas av rättigheten.[26] Det som avses med skapande är den faktiska insatsen som leder fram till att *verket* blir till.

Den ekonomiska ensamrätten, eller förfoganderätten om man så vill, innebär att upphovspersonen bland annat har uteslutande rätt att framställa exemplar av verket, jfr 1 kap. 2 § upphovsrättslagen.[27] Att upphovsrätten tillkommer upphovspersonen ensam innebär att han eller hon bland annat kan begränsa andras möjligheter att använda, mångfaldig-

[24] Undantaget i 1 kap. 9 § upphovsrättslagen att upphovsrätt inte uppkommer vid upprättande av vissa handlingar hos myndighet träffar normalt inte den sorts material som skapas av enskilda lärare. Materialet kan dock tänkas utgöra allmän handling, vilket leder till att det utan hinder av upphovsrätten kan lämnas ut med stöd av 2 kap. tryckfrihetsförordningen, jfr 2 kap. 26 b § upphovsrättslagen.

[25] Se Olsson & Rosén, Upphovsrättslagstiftningen, JUNO version 4A 2018 [Olsson & Rosén (2018)] avsnittet om Upphovsrätten och de närstående rättigheterna, Bernitz m.fl. (2020) s. 69 ff.

[26] Se Olsson & Rosén (2018) avsnittet om Upphovsrätten och de närstående rättigheterna.

[27] Regleringen är harmoniserat inom EU genom direktiv 2001/29/EG (Infosoc).

göra och sprida materialet. Skyddsomfånget beror emellertid på graden av originalitet. Verk som bara når upp till den nedre tröskeln för verksbegreppet skyddas bara mot rena efterbildningar, medan upphovsrättsintrång kan ske även genom alster som uppvisar skillnader mot originalverket om detta utvisar högre grad av originalitet.[28]

Upphovspersonen kan givetvis samtycka till att andra ska få använda dennes verk, och med de begränsningar som följer av det ideella skyddet i 1 kap. 3 § upphovsrättslagen kan upphovsrätten också överlåtas.[29] Det finns också vissa undantag i 2 kap. upphovsrättslagen som på olika sätt innebär inskränkningar i upphovspersonens ekonomiska ensamrätt. Dessa regler löser emellertid bara i mycket begränsad utsträckning de upphovsrättsliga frågor som kan uppkomma i relationen mellan läraren som arbetstagare och lärosätet som arbetsgivare gällande rätten att nyttja upphovsrättsskyddat undervisningsmaterial.[30]

Dessa upphovsrättsliga utgångspunkter gäller generellt, och det spelar i detta sammanhang ingen principiell roll om upphovspersonen har skapat verket inom ramen för en lärartjänst eller i annat sammanhang.[31]

2.3 Möjliga inskränkningar i lärarens ekonomiska ensamrätt till följd av anställningsförhållandet: Tumregel och lärarundantag

Fastän läraren har upphovsrätt till sina verk, kan vissa inskränkningar i den ekonomiska ensamrätten tänkas föreligga som innebär att lärosätet kan nyttja materialet i sin verksamhet. Det kan för sammanhangets skull

[28] Se Sandgren (2003) s. 22.

[29] Se 3 kap. 27 § och de i 3 kap. 28 § aktuella begränsningarna i förvärvarens rätt att förfoga över förvärvad upphovsrätt.

[30] Citaträtten och möjligheten enligt 3 kap. 42 c § upphovsrättslagen att enligt kopieringsavtal framställa exemplar av delar av publicerade texter till bruk i undervisningen, t.ex. i materialsamlingar, är givetvis av stor betydelse för verksamheten, men upphovspersonen kan enligt 3 kap. 42 c § andra stycket meddela förbud mot exemplarframställningen. Se närmare om avtalslicenser inom upphovsrätten Lund, Den nordiske planten avtalelisens. Forgreninger eller avleggere …, Liber Amicorum festskrift till Jan Rosén s. 505 ff. Rätten att använda texter i enlighet med avtalslicensen på området är också oberoende av om materialet skapats och publicerats av en av lärosätets egna lärare, eller någon helt utomstående. En mycket praktiskt viktig begränsning i en annan, men för universitetsläraren angränsande, fråga av upphovspersonens ensamrätt finns i 2 kap. 12 § upphovsrättslagen om exemplarframställning till privat bruk.

[31] Se t.ex. Karnell (1969) på s. 54.

vara lämpligt att redan här introducera två okodifierade regler som tenderar att dyka upp i nära anslutning till frågeställningen, nämligen *tumregeln och lärarundantaget*.

Inom en del av patenträtten, nämligen den del som gäller rätten till arbetstagares uppfinningar, har det länge funnits ett lagstadgat *lärarundantag* som anses gälla även upphovsrättsliga frågor.[32] Det patenträttsliga lärarundantaget innebär att lärare vid universitet, högskolor eller andra inrättningar som tillhör undervisningsväsendet *inte* ska anses såsom arbetstagare enligt lagen om rätten till arbetstagares uppfinningar.[33] Detta betyder att de i nämnda lag uppställda *inskränkningarna* av den annars gällande utgångspunkten att arbetstagare – som andra uppfinnare

[32] Se 1 § lagen (1949:345) om rätten till arbetstagares uppfinningar (LAU) och t.ex. SOU 2020:59 Innovation som drivkraft – från forskning till nytta s. 67. Att det finns även ett upphovsrättsligt lärarundantag tycks de flesta vara överens om, och nästan oavsett vilket håll man ser åt hittar man påståenden om att ett sådant finns sedan länge. Se t.ex. SOU 2010:24 Avtalad upphovsrätt s. 164. I SOU 2020:59 s. 67 anges att avgränsningen som följer av LAU endast finns i teorin, men att lärarundantaget i praktiken används ”som en sammanfattande benämning på de högskoleanställda lärarnas rätt till resultat de i vid mening är upphovsmän till i sin anställning.” I Lund, Immaterialretten og forskningen ved universiteter og høyskoler i NIR 2000 s. 618–629 [Lund (2000)] framhålls på s. 621 att det är allmänt accepterat att den okodifierade regeln om övergång av upphovsrätt till följd av ett anställningsförhållande inte gäller universitetslärares forskning och heller inte deras undervisningsmaterial, dock utan hänvisning till några bestämda källor. Schaumburg-Müller, Videnskabeligt personales ophavsretlige stilling, NIR 1986 s. 282–291 [Schaumburg-Müller (1986)] på s. 286 lade också uttryckligen till grund att det vid tidpunkten följde av fast dansk teori och praxis att anställningsförhållandet som sådant inte innebar någon övergång av upphovsrätt från universitetslärare till lärosätena. Se också Strömholm (2002) s. 41 som anger att ingen övergång av upphovsrätt sker gällande lärarens forskningsinsatser. Se också Kielland (2018) s. 49. Att hitta auktoritativa rättskällor som stödjer existensen av ett upphovsrättsligt lärarundantag är emellertid mycket svårt.

[33] Det patenträttsliga lärarundantaget torde vara en av de regler som ägnats mest uppmärksamhet i offentliga utredningar de senaste 30 åren, dock utan att nödvändigtvis ha lett till lagstiftning. Se t.ex. SOU 1944:27, Rätten till vissa uppfinningar m.m, SOU 1977:63, Fortsatt högskoleutbildning, SOU 1980:42, Arbetstagares uppfinningar, SOU 1996:29, Forskning och pengar, SOU 1996:70 (NYFOR), Samverkan mellan högskolan och näringslivet, prop. 1996/97:5, Forskning och samhälle, prop. 1998/99:94, Vissa forskningsfrågor, SOU 1998:128, Forskningspolitik, prop. 2000/01:3, Forskning och förnyelse, VP2003:1, Vinnforsk, prop. 2004/05:80, Forskning för ett bättre liv, SOU 2005:95 Nyttiggörande av högskoleuppfinningar och SOU 2020:59 Innovation som drivkraft – från forskning till nytta. I flera av dessa, som i SOU 2010:24 Avtalad upphovsrätt, finns också spridda kommentarer om ett upphovsrättsligt lärarundantag.

– har rätt till sina uppfinningar, inte gäller för denna yrkesgrupp. För lärare gäller alltså principen att uppfinnaren har rätt till sina uppfinningar *oinskränkt*.[34]

Något motsvarande undantag finns inte i upphovsrättslagen, men motsvarande regel har alltså på sedvanerättslig grund ansetts föreligga också inom upphovsrätten, om än med oklara konturer och osäker räckvidd.[35] Mot bakgrund av den patenträttsliga regeln antyder termen *lärarundantag* att läraren till följd av undantaget får en mer oinskränkt rättighet än den som annars hade förelegat.[36] Förhållandet mellan upphovsrättslagen och ett eventuellt lärarundantag är emellertid inte helt okomplicerat, och det är inte uppenbart att rättsläget är helt jämförbart med vad som gäller avseende arbetstagares uppfinningar.[37] Upphovsrätten uppkommer trots allt enligt gällande rätt per automatik, och tillfaller den som *skapat* verket. Detta i kombination med att det *inte finns* motsvarande lagstadgade allmänna inskränkningar av upphovsrätten som de som gäller för arbetstagares uppfinningar,[38] gör att det kan ifrågasättas om det alls finns något behov av ett lärarundantag inom upphovsrätten och vad ett sådant i så fall innebär.

Frågan om ett upphovsrättsligt lärarundantag hänger emellertid nära samman med den andra okodifierade regeln, nämligen den så kallade

[34] Se om lärarundantaget främst ur patenträttslig, men även upphovsrättslig, synvinkel, Wolk (2008) s. 205 ff. De hänsyn som det patenträttsliga lärarundantaget bygger på gör sig, som Sandgren (2003) s. 38 framhåller, än starkare gällande när tematiken i stället är lärarens upphovsrätt.

[35] Se Wolk (2008) s. 210 f., Sandgren (2003) s. 38 ff. och Strömholm (2002) s. 41. I Finland infördes emellertid 1991 ett lärarundantag kopplat till bestämmelsen om övergång av upphovsrätt till datorprogram skapade i tjänsten, jfr 3 kap. § 40 b 2 st. finska upphovsrättslagen. Se närmare Bruun, Upphovsrätt i anställning – nuläge och utvecklingstendenser, Vennebog till Mogens Koktvedgaard, red. Marianne Levin, (1993) s. 152–167 [Bruun (1993] på s. 154.

[36] Som påpekats i SOU 2005:95 torde det emellertid inte få någon större betydelse för lärarens rättigheter om det patenträttsliga lärarundantaget skulle tas bort. Se särskilt resonemanget på s. 175 med vidare hänvisningar till SOU 1996: 70 (NYFOR) där motsvarande slutsatser framgår.

[37] En poäng som gjordes också av Schaumburg-Müller (1986) s. 283.

[38] I 3 kap. 40 a § upphovsrättslagen finns dock ett specifikt motsvarande undantag, som uttryckligen anger att ”[u]pphovsrätten till ett datorprogram, som skapas av en arbetstagare som ett led i hans arbetsuppgifter eller efter instruktioner av arbetsgivaren, övergår till arbetsgivaren, såvida inte något annat har avtalats.” Bestämmelsens räckvidd och tillämpning på högskolelärare har dock diskuterats. Se Bruun (1993) s. 158 ff. för en översikt och vidare hänvisningar.

tumregeln.[39] Tumregeln innebär att en arbetsgivare på närmare angivet sätt, och då *uteslutande till följd av anställningsavtalet*, äger rätt att nyttja och vid behov ändra och anpassa upphovsrättsskyddat material som har skapats av en anställd.[40] Också den upphovsrättsliga tumregelns konturer är otydliga, och regelns räckvidd är inte i alla delar lätt att avgöra. Detta har ett naturligt samband med att regelns tillämpning skiljer sig åt både mellan olika branscher och mellan olika personalkategorier inom samma bransch, men även på samma arbetsplats. Arbetsgivarens nyttjande kan nämligen med stöd i denna regel ske i *den mån som det behövs* för att anställningsavtalet ska fylla sitt syfte och som led i arbetsgivarens vanliga verksamhet vid tidpunkten för verkets tillkomst.[41] Detta innebär att det måste bedömas konkret för olika anställningsförhållanden huruvida någon övergång alls sker, och i så fall i vilken omfattning.[42]

Ett lärarundantag inom upphovsrätten skulle mot denna bakgrund kunna tänkas utgöra ett undantag från tumregeln som en okodifierad

[39] Tumregeln som regel formulerades så långt jag kan se först av Karnell, och beskrivs i Karnell, Läromedelsrätt, 1972 [Karnell (1972)] s. 35. I domen AD 2002 nr 87 har Arbetsdomstolen formulerat regeln så att ”[e]n arbetsgivare får inom sitt verksamhetsområde och för sin normala verksamhet utnyttja sådana verk som tillkommer som ett resultat av tjänsteåligganden gentemot arbetsgivaren. Arbetsgivarens rätt avser utnyttjanden för de ändamål som kan förutses när verket tillkommer.”

[40] Se SOU 2010:24 Avtalad upphovsrätt s. 144 ff. för en genomgång av doktrin och praxis på området. Se också i Kielland (2018) s. 45 ff. med vidare hänvisningar till ytterligare nordiska källor. Som tumregeln formulerats i doktrinen, innebär den inte endast en rätt att nyttja upphovsrättsskyddat material i verksamheten, utan även att göra ändringar i materialet utan upphovspersonens samtycke. Sådana ändringar måste emellertid ligga inom ramen för 1 kap. 3 § upphovsrättslagen. Se Wolk, Anställdas immaterialrätter – divergenta förhållanden, Ny Juridik 4:04 s. 17 ff. på s. 20.

[41] I allmänna framställningar om arbetstagares upphovsrätt finns olika uppfattningar om hur långtgående arbetsgivarens rätt att nyttja de upphovsrättsskyddade verken skapade av en anställd är. De två huvudsakliga alternativen är om (i) bedömningen ska vara differentierad och utgå från omständigheterna i det enskilda fallet utifrån vad som behövs, eller att (ii) arbetsgivaren ska anses få rätt till allt som skapas inom ramen av tjänsten. Som redan Godenhielm (1978) s. 331 summerar, vill emellertid arbetsgivaren oavsett få en nyttjanderätt till verket ”i den omfattning som följer av syftet med anställningsavtalet”. Se Svensäter, Anställning och upphovsrätt, 1991 [Svensäter (1991)] s. 332–350 för en närmare beskrivning av olika författares ståndpunkt i frågan.

[42] Motsvarande bedömning av tumregelns räckvidd måste göras om det istället handlar om en uppdragstagare som skapar verk på beställning. Se Wolk, Hyrda arbetstagares upphovsrätt, NIR 2001 s. 210–217 [Wolk (2001)] på s. 216, Calissendorff, Rätten till ett beställt verk, Ny Juridik 2:98 s. 89 ff. och Bengtsson, Rätten till beställt verk i ljuset av nya domstolsavgöranden, Ny Juridik 4:05 s. 31 ff.

lära om rätt att nyttja upphovsrättsskyddat material i anställningsförhållanden.[43] Ett *absolut* lärarundantag skulle då ha som effekt att *ingen* övergång av upphovsrätt till material som har skapats av en lärare sker till följd av anställningsförhållandet, och att någon konkret bedömning av vad som med stöd i tumregeln eventuellt hade gått över till arbetsgivaren blir onödig. Stöd för ett så absolut lärarundantag är emellertid svårt att finna. I realiteten framstår snarare regelns räckvidd som otydlig och beroende på vilken sorts material det handlar om.[44]

Frågan om lärosätets rätt att nyttja det av läraren skapade, upphovsrättsskyddade, materialet, är – oavsett hur förhållandet mellan tumregeln och lärarundantaget ska hanteras – en fråga om upplåtelse eller överlåtelse av upphovsrätt som innebär en viss inskränkning av upphovspersonens ekonomiska ensamrätt. Frågan om lärosätets nyttjanderätt är däremot *inte* av betydelse för om läraren över huvud taget har upphovsrättsligt skydd för sina verk eller för lärarens möjlighet att själv offentliggöra eller framställa exemplar i andra sammanhang.[45]

[43] Wolk tycks 2011 vara av den uppfattningen att det krävs uttryckligt tillstånd från läraren om ett lärosäte över huvud taget ska kunna nyttja lärarens utbildningsmaterial och som stöd för detta anförs uteslutande lärarundantaget. Se Wolk, Universitetslärarens upphovsrätt, XXXVIII SULF:s Skriftserie 2011 [Wolk (2011)] s. 17. Wolks rapport är emellertid mycket kortfattad och varken källor eller bakomliggande resonemang redovisas. Eftersom synpunkterna som framgår i rapporten också i någon mån avviker från vad som följer av hennes avhandling, Wolk (2008) s. 205 ff., är det svårt att tillmäta rapporten något värde som rättskälla.

[44] Se Sandgren (2003) s. 39 f. och Strömholm (2002) s. 41. Även Wolk (2008) s. 211 framhåller att det kan vara svårt att ange exakt var gränserna för sedvanan går eftersom olika policy och avtal finns vid olika högskolor. Motsvarande synpunkt framgår i arbetsrättslig doktrin, t.ex. Källström & Malmberg, Anställningsförhållandet: inledning till den individuella arbetsrätten, 4 uppl., 2016 [Källström & Malmberg (2016)] s. 240 f.

[45] Det kan givetvis tänkas föreligga arbetsrättsliga begränsningar för hur läraren kan agera och utnyttja sitt material, t.ex. i konkurrerande eller förtroendeskadande verksamhet eller liknande, men det är alltså begränsningar som inte har med den upphovsrättsliga frågeställningen att göra. Se lite närmare om gränsytan mellan upphovsrätt och arbetsrätt i detta sammanhang i Wolk, Hyrda arbetstagares upphovsrätt, NIR 2001 s. 210–217 på s. 212.

3 Övergång av upphovsrätt i universitetslärarens anställningsförhållande

3.1 (Underförstått) avtal eller tumregel?

Lärosätets eventuella rätt att nyttja det undervisningsmaterial som skapats av dess lärare är som visat ovan en från upphovspersonen härledd rättighet. Den principiella utgångspunkten är därmed att lärosätets nyttjande behöver någon form av rättsligt stöd. Om tumregeln ger sådant stöd i dessa fall, eller om avtal med läraren behövs, kan emellertid framstå som osäkert.[46]

Om tumregeln är en sedvanerättslig regel som gäller som dispositiv bakgrundsrätt i fall då inget annat avtalats mellan parterna, eller om tumregeln snarare ger uttryck för en *tolkningsprincip* om vad som i sådana fall kan presumeras följa av ett underförstått avtal mellan parterna, är inte i alla delar tydligt.[47] Tumregeln kan nog historiskt sett anses härledd ur just sedvänja eller branschpraxis som grundar sig i synpunkter om underförstådda samtycken i de anställningsförhållanden då sådan rättighetsövergång behövs för förverkligandet av anställningsförhållandets syfte, men hur den ska förstås i dag är inte i alla delar tydligt. Ett avtal kan också vara skriftligt eller muntligt, uttryckligt eller underförstått, och det är särskilt i anknytning till de underförstådda avtalen som kopplingen till tumregeln framträder tydligast, och där gränsen mellan tumregeln och ett underförstått avtal blir som mest svårdragen.[48]

I de flesta fall kan man nog utgå ifrån att det föreligger ett stillatigande samtycke från upphovspersonen, så att arbetsgivaren faktiskt har det samtycke som behövs för sitt nyttjande av det upphovsrättsskyddade materialet,[49] åtminstone så länge som den anställde inte motsätter sig

[46] Se Karnell (1972) s. 36 som framhåller att det bör träffas skriftliga avtal i det enskilda fall eftersom ”[d]et är alltför osäkert, för alla parter, att förlita sig på de allmänna regler som eljest gäller.”

[47] I SOU 2010:24 s. 163 tycks den slutsats som dras mot bakgrund av uttalanden i doktrinen vara att det handlar om en presumtion om övergång då konkreta hållpunkter saknas, vilket får förstås som att regeln utgör sedvanerätt som i någon mening behöver avtalas bort.

[48] Se t.ex. Schaumburg-Müller (1986) s. 285 och Kielland (2018) s. 41.

[49] Huvudregeln avtalsrättsligt är trots allt att det inte gäller formkrav, och att en viljeförklaring kan vara uttrycklig eller underförstådd, eller rentav härledas ur ett konkludent handlande eller en underlåtenhet att handla (passivitet). Se t.ex. Adlercreutz, Avtalsrätt I, 14 uppl. [Adlercreutz (2016)] s. 60. Allmänna avtalsrättsliga principer gäller också på

den från arbetsgivaren önskade användningen, fastän han eller hon vet om det aktuella bruket.[50] I normalfallet kommer mycket av det material som läraren skapar till användning på en kurs att hamna i denna kategori, där ett samtycke kan anses ha getts och där läraren inte heller har något intresse av att begränsa lärosätets användning. Om arbetsgivarens nyttjanderätt bygger på ett underförstått samtycke från arbetstagaren, innebär det visserligen att svåra tolkningsfrågor om samtyckets räckvidd kan uppkomma,[51] men också att läraren vid skapande av nya sorters material, eller då arbetsgivaren vill nyttja ett material på ett nytt sätt, *kan* motsätta sig arbetsgivarens önskade användning.[52] I normalfallet kan man dock inte förvänta sig att läraren ska agera helt som en vilken som helst annan mer fristående avtalspart. Att i en löpande anställning protestera mot arbetsgivarens användning av det material läraren skapat, skulle lätt kunna tänkas ge negativa arbetsrättsliga effekter i form av lönepåverkan och påstådda samarbetsproblem. I kombination med att arbetstagaren står under arbetsledning och är skyldig att följa arbetsgivarens order även under tvist, innebär de arbetsrättsliga särdragen att man nog får vara försiktig med att tolka in för mycket i en eventuell passivitet, åtminstone så länge som anställningen består.

det upphovsrättsliga området, om än i kombination med specifikationsprincipen som innebär att ”rättighetshavaren inte antas ha upplåtit en mera omfattande rätt än vad som klart framgår”, jfr Bernitz m.fl. (2020) s. 384 f. Se också Rt. 2001 s. 872 och Kielland (2018) s. 42.

[50] Att upphovspersonen vet om hur materialet används, eller åtminstone bör ha sådan kunskap, är en förutsättning för att dennes underlåtenhet att agera (passivitet) ska kunna anses som uttryck för en vilja och således utgöra en grund för bindning. Se närmare Vea Lund, Passivitet, 2017, s. 189 ff. om löftegivarens kunskap som allmän förutsättning för bundenhet genom passivitet.

Se också Adlercreutz (2016) s. 117 ff.

[51] Att tolka underförstådda viljeförklaringar i en kontext då praktiken vid olika lärosäten, och även inom den enskilda institutionen, kan variera och sällan kommer tydligt till uttryck, och där olika personer kan ha mycket olika uppfattningar kring vad som gäller, torde helt enkelt vara relativt omöjligt. Otydlighet eller oklarhet kommer emellertid i dessa typfall, mot bakgrund av specifikationsprincipen, lösas till förmån för upphovspersonen. Vid oklarhet ska överlåtelsen tolkas restriktivt. Se t.ex. Godenhielm (1978) s. 323, Bernitz m.fl. (2020) s. 385 och Kielland (2018) s. 42.

[52] Riktlinjer uppställda av arbetsgivaren har i sammanhanget ingen självständig eller principiell rättslig betydelse för frågan om övergång av upphovsrätt i anställningsförhållandet. Utgångspunkten är helt enkelt att arbetsgivaren inte genom att uppställa riktlinjer kan ingripa i den anställdes civilrättsliga positioner som annars förutsätter avtal eller annan rättsgrund.

Som *sedvanerättslig* regel skulle tumregeln kunna innebära att vissa upphovsrättsliga rättigheter övergår från universitetsläraren till lärosätet helt *oberoende* av den anställdes vilja i det enskilda fallet, om inte annat avtalats konkret. Det är därför inte nödvändigtvis betydelselöst om det handlar om tolkning av ett underförstått avtal eller en sedvanerättslig regel, förutsatt att övergång av upphovsrätt *överhuvudtaget* kan tänkas ske med stöd av tumregeln i den aktuella situationen. När det gäller rätten till undervisningsmaterial som har skapats av lärare på högskola och universitet är emellertid den enligt min uppfattning mer angelägna frågan huruvida tillämpning av tumregeln *överhuvudtaget* ska leda till övergång av upphovsrätt till lärosätena som arbetsgivare.

Det kan inte uteslutas att lärosätena framöver börjar reglera upphovsrättsliga frågor i sina anställningsavtal. Avtalsfrihet råder, åtminstone avseende förfoganderätten, och huvudregeln enligt 3 kap. 27 § upphovsrättslagen är att upphovspersonen kan överlåta de ekonomiska rättigheterna som följer av upphovsrätten.[53] Det torde emellertid kunna ifrågasättas hur reell avtalsfriheten är om arbetsgivaren uppställer som villkor för en nyanställning att övergång av upphovsrätt till verk skapat av läraren ska ske. Det är stor konkurrens om tjänster och de flesta nydisputerade skulle sannolikt acceptera nästan vilka villkor som helst för möjligheten att meritera sig, eller även få en fast tjänst. Denna sorts avtalsvillkor skulle också principiellt vara svåra att få ihop med 2 kap. 16 § RF och regleringen av den akademiska friheten, jfr 1 kap. 6 § 2 st. 3 högskolelagen.[54] Om överlåtelsen går utöver vad som egentligen behövs och vad som har stöd i rättskällorna, och vad som accepterats av facket, torde jämkning med stöd av 36 § avtalslagen ligga nära till hands.

[53] I sådana fall måste man dock skilja mellan nyanställda och de sedan tidigare anställda, jfr också Wolk (2008) s. 213. Det skulle utgöra en markant ändring av villkoren för anställningsavtalet att införa överlåtelse av upphovsrätten som ett krav i den enskildes redan existerande arbetskontrakt. En sådan ändring faller utanför arbetsgivarprerogativens domän, och är därmed inte något som kan genomföras av arbetsgivaren utan arbetstagarens samtycke, jfr Glavå & Hansson, Arbetsrätt, 3 uppl., 2016 [Glavå & Hansson (2016)] s. 573 f. om § 32-befogenheterna och dess begränsningar.

[54] Bestämmelsen i högskolelagen anger att forskningsresultat får publiceras fritt, vilket rimligen måste betyda att författaren, alltså forskaren, äger att välja hur och när publicering ska ske utan inblandning från andra. Detsamma följer av upphovsrätten, enligt vilken författaren har en oinskränkt rätt till sina verk innan de har offentliggjorts. Se Wolk (2009) s. 723.

3.2 Anställningsförhållandets särprägel och kopplingen till arbetsrätten

3.2.1 Inledning

I grunden bygger tumregeln på en bedömning av vilken övergång av upphovsrätt från den anställde till arbetsgivaren som är *rimlig och nödvändig* för att verksamheten ska kunna fungera på det vis som förutsattes vid anställningsavtalets ingående, med andra ord en bedömning av vad som krävs för att anställningsförhållandets syfte ska uppnås.[55] Regeln hänger nära samman med att den som anställts för att skapa upphovsrättsligt skyddade verk genom den månatliga löneutbetalningen får betalt för att arbetsgivaren ska kunna nyttja materialet som förutsatt.[56] Mot denna bakgrund behöver emellertid något ytterligare sägas om vad som är särpräglat med universitetslärarens anställning.[57]

Inom högskolesektorn kan en tumregel tänkas vara av mycket stor praktisk betydelse, eftersom det är sällsynt att lärare och lärosäte genom uttryckliga avtal reglerar frågor om rätten till användning av upphovsrättsskyddat material skapat inom ramen för vanliga anställningar som lärare.[58] Samtidigt är det inte så att existensen av en tumregel på området skulle leda till att lärosätet får nyttjanderätt till *allt* upphovsrättsskyddat material skapat av en lärare, utan endast till vad som är nödvändigt för lärosätets verksamhet och relevant för anställningsavtalets syfte. Som anges i SOU 2010:24 s. 164 har anställningsförhållanden inom forsk-

[55] Se Karnell (1972) s. 34 och Godenhielm (1978) s. 329. Övergång av upphovsrätt i anställningsförhållanden följer alltså inte den annars gällande huvudregeln att äganderätten till det som produceras i en verksamhet av dess anställda tillfaller arbetsgivaren. Se närmare Wolk (2008) s. 20 ff.

[56] Se Karnell (1972) s. 33–34 om att en grundläggande förutsättning för att frågan om upphovsrättens övergång ska uppkomma är att verket framställts som ett tjänsteåliggande eller på grund av ett särskilt åtagande gentemot arbetsgivaren. Se också Godenhielm (1978) s. 329.

[57] Se Karnell (1972) s. 36–37 om tumregelns (begränsade) utbyte för arbetsgivare på universitets- och högskoleutbildningens område, åtminstone när det gäller forskning och läromedel. Se också Wolk (2008) kap. VII om särskilda förhållanden vid universitet och högskola.

[58] Det vanliga torde i stället vara att läraren skapar och utvecklar det material som behövs för att en kurs ska fungera utan att nödvändigtvis tänka på de upphovsrättsliga frågeställningar som kan aktualiseras. Hållningen är gärna att undervisningsmaterialet skapas ”för kursen” och är avsett att återanvändas, om än under vissa, gärna underförstådda, förutsättningar, som kan variera mellan olika lärare.

nings- och utbildningsområdena normalt *inte* ansetts motivera något rättsförvärv för arbetsgivaren gällande vetenskaplig produktion eller läroböcker.[59] Det är främst undervisningsmaterial och annat material som behövs för genomförandet av lärosätets faktiska verksamhet som potentiellt kan tänkas omfattas av en upphovsrättslig tumregel.

Inom anställningsförhållanden där den anställdes jobb främst är att skapa upphovsrättsskyddat material för användning i arbetsgivarens verksamhet vore det egendomligt om den anställde inte också var inställd på att låta arbetsgivaren nyttja det aktuella materialet på det sätt som tjänsten förutsätter. Att motsätta sig sådan användning skulle möjligen kunna genomdrivas rent upphovsrättsligt, men hade samtidigt inneburit en allvarlig brist i den anställdes skyldighet att fullgöra sina förpliktelser enligt anställningsavtalet, vilket man får förmoda att den anställde inte anser sig betjänt med.[60] Det är i denna sorts typfall som tumregeln har ansetts få störst betydelse och där kollektivavtal gärna omfattar frågor om upphovsrättens övergång. Typiskt sett uppkommer tvistefrågor i dessa fall först då arbetsgivaren använt materialet på *annat* sätt än som ursprungligen förutsattes.[61]

En sådan tydlig spänning mellan de upphovsrättsliga och arbetsrättsliga utgångspunkterna kommer emellertid inte gärna uppstå för den klassiska universitetsläraren.[62] Lärarens tjänst innefattar nämligen en rad olika uppgifter, där de som utgör den största delen av arbetstiden inte alls *behöver* innefatta ett skapande av upphovsrättsskyddat material,[63] och där andra uppgifter leder till skapande av upphovsrättsligt skyddat material som lärosätena vare sig behöver, eller kommer få, rätt att nyttja med stöd av tumregeln.[64] Den enskilde läraren, åtminstone på större utbildningar, skapar exempelvis normalt sett inget eget material inför sin seminarie-

[59] Ett eventuellt upphovsrättsligt lärarundantag får också främst anses gälla just denna sorts material, jfr Strömholm (2002) s. 80.

[60] Se Glavå & Hansson (2016) s. 505 och Källström & Malmberg (2016) s. 217.

[61] Se t.ex. NJA 1999 s. 390, NJA 2004 s. 363, RH 2009:7, AD 2002 nr 87, U 1978.901 H.

[62] Se Kielland (2018) s. 50.

[63] Av 3 kap. 1 § högskolelagen följer ingen skyldighet för läraren att skapa sådant material alls, men många kommer av pedagogiska skäl att göra det i samband med fullgörelsen av delar av sin undervisningsskyldighet fastän det inte kan krävas av arbetsgivaren, jfr närmare nedan under 3.2.2. Se också Wolk (2008) s. 212 f.

[64] Det tydligaste exemplet i denna kategori torde vara den anställdes vetenskapliga produktion, men också läromedel faller tydligt utanför lärosätets nyttjanderätt.

undervisning, utan undervisar snarare år efter år med utgångspunkt i befintliga uppgifter, och samhällsuppdraget fullgörs genom att forskarens vetenskapliga alster och eventuella läroböcker och forskning publiceras och på så sätt blir tillgängliga för omvärlden.[65]

Vad som gäller när läraren trots allt har skapat upphovsrättsskyddat undervisningsmaterial är givetvis av stort intresse, och ska behandlad närmare nedan, men det kan ändock vara bra att ha med sig att sådant skapande inte nödvändigtvis tar i anspråk någon större del av lärarens arbetstid. Det *behöver* alltså inte uppkomma någon konflikt mellan den arbetsrättsliga och den upphovsrättsliga regleringen gällande universitetslärarens verk, och om konflikten ens uppstår kan den mycket väl gälla material skapat på en helt begränsad del av lärarens totala arbetstid. Detta leder till att vi, som jag ser det, redan till en början är i ytterkanten av tumregelns tillämpningsområde.

3.2.2 Akademisk frihet och arbetsgivarens (mycket begränsade) rätt att leda och fördela arbetet

Eftersom tumregeln i grunden bygger på kopplingen mellan anställningsförhållandet och den anställdes skapande som ett led i fullgörandet av sin arbetsskyldighet,[66] är det av betydelse att se närmare på lärarens eventuella skyldighet att skapa undervisningsmaterial. Rollen som universitetslärare är trots allt inte i alla delar lätt att jämföra med jobb där arbetsgivarens rätt att leda och fördela arbetet också gäller *hur* olika uppgifter ska utföras. I någon mån kunde det som ofta sammanförs under beteckningen akademisk frihet sägas stå i viss motsättning till arbetsledningsrätten. Även denna motsättning är av betydelse för frågan om övergång av upphovsrätt från lärare till lärosäte till följd av anställningen.[67]

[65] Ofta är det lärosätenas verksamhet i relation till exploatering och kommersialisering av upphovsrättsskyddat material som problematiseras, jfr t.ex. Wolk (2008) s. 206, men också återanvändning i den faktiska verksamheten är av intresse här.

[66] Arbetsdomstolen formulerade i AD 2002 nr 87 detta som ett krav om att förfoganderätten gäller nyttjande av verk som tillkommer som resultat av tjänsteåligganden gentemot arbetsgivaren. Se också Karnell (1972) s. 33, Godenhielm (1978) s. 327 ff. och Strömholm (2002) s. 79 f.

[67] Se Karnell (1972) s. 35 f. om att arbetsgivarens anvisningar om ett verks utformning är utan upphovsrättslig betydelse om framställningen ligger utanför den anställdes tjänsteanliggande, och att det kan bli nödvändigt att fastställa om den som givit anvisningar faktiskt varit (arbetsrättsligt) behörig att ge direktiv.

Arbetsgivarens ledningsrätt kan i typfallet med universitetsläraren sägas vara närmast obefintlig vad gäller den anställdes forskning.[68] Både vad som görs till forskningsämne, val av forskningsinriktning och genomförande,[69] samt till stor del när forskningen bedrivs, är den enskildes val.[70] Detsamma gäller vid produktion av läromedel.[71] Vad gäller undervisning kan lärosätet och institutionen givetvis fastställa styrdokument, vilka i viss mån kommer att begränsa lärarens frihet. Lärosätet kan givetvis också ålägga den enskilde läraren olika undervisningsuppgifter.[72] Någon ledningsrätt avseende innehållet eller hur läraren väljer att genomföra sin undervisning föreligger däremot inte.[73] Avseende administrativa frågor, såsom till exempel schemaläggning, kursansvar och resurstilldelning för olika kurser, gäller en mer normal och omfattande arbetsledningsrätt, givetvis begränsad av anställningsavtalet samt tjänstens grundprägel.[74]

Lärarens förpliktelser avseende den sorts uppgifter som räknas till undervisningsdelen av tjänsten kan ses dels som en skyldighet att frambringa ett bestämt resultat (dyka upp till sin schemalagda undervisning, skapa en tentamensuppgift, kommentera ett visst antal PM, skriva en lärarhandledning till ett seminarium *et cetera*), dels som en skyldighet att göra så gott den kan i att framställa dessa resultat. Bedömningen av vad som är bra eller dåligt avseende innehållet eller sättet att utföra uppgiften, och vilken form eller utformning resultatet ska ha, görs däremot till stor del av den enskilde läraren. Rådande för lärarens arbete i detta avseende är den akademiska friheten vilken innebär att läraren som utgångspunkt själv, om än inom ramen för gällande reglering av den del av uppgifterna

[68] Så också Bruun (1993) s. 158 om kopplingen till arbetstagarens ”uppgifter” i samband med frågan om övergång av upphovsrätt till datorprogram i anställningsförhållanden.

[69] Jfr särskilt 1 kap. 6 § högskolelagen och 2 kap. 16 § RF.

[70] De flesta lärare har förtroendearbetstid vilket innebär att man utöver schemalagd undervisning och eventuell kontorstid normalt kan utföra sitt jobb när på dygnet som man helst vill, om man bara passar på att göra tillräckligt många timmar.

[71] Se Kielland (2018) s. 50.

[72] När arbetsgivaren utöver sin arbetsledningsrätt är det dock viktigt att såväl god sed i anställningsförhållanden som likabehandlingsprincip och saklighets- och opartiskhetskrav som gäller i den offentliga sektorn iakttas. Se Svensäter (1991) s. 92.

[73] Se Wolk (2008) s. 205 f.

[74] Se Glavå & Hansson (2016) s. 585 ff. om arbetsledningsrätten och dess begränsningar.

som utgör myndighetsutövning, kursplan eller andra formella styrdokument, har frihet att välja både hur och när uppgifterna ska genomföras.[75]

Eftersom läraren som huvudregel själv väljer *hur* den vill fullgöra de uppgifter den har åtagit sig eller blivit ålagd, och därmed också vilken sorts material som eventuellt ska skapas i samband med detta, måste det också vara den enskilda lärarens sak att välja om materialet alls ska användas, och om det ska återanvändas, ändras eller helt plockas bort inför kommande kursomgångar.[76] Fastän läraren exempelvis hållit samma föreläsningar år efter år och har skapat förinspelat material för att ersätta föreläsningarna, behöver fullgörelsen av undervisningen inte alls ske genom att detta material återanvänds, utan kan likväl ske genom att läraren i stället ställer sig upp på utsatt tid och håller sin föreläsning på plats i en hörsal. Att arbetsgivaren i ett sådant fall inte med stöd i arbetsledningsrätten kan ålägga läraren att i stället använda det förinspelade materialet är uppenbart, precis som att arbetsgivaren inte kan ålägga läraren att spela in materialet från första början.[77]

Något svårare kan frågan om arbetsledningsrätten, och kopplingen mellan tjänstens innehåll och skapandet av upphovsrättsskyddat material, bli när det gäller material som skapas mer eller mindre på direkt beställning från arbetsgivaren, som till exempel tentamensuppgifter eller seminarieuppgifter.[78] Det anses normalt ingå i tjänsten som lärare att man ”tar sin andel” av examinationen, vilken förutsätter upprättande av tentamensuppgifter, såväl som att man tar del av utveckling och uppdatering av undervisningsmaterial. Det kan dock variera mellan olika

[75] Just den akademiska friheten har också framhållits som en central bakgrund till varför någon övergång av upphovsrätt i anställningsförhållanden inte torde vara aktuellt. Se SOU 1996:70 Samverkan mellan högskolan och näringslivet s. 92.

[76] Se Wolk (2008) s. 212 f. Förhållandet mellan arbetsgivarens arbetsledningsrätt och den anställdes lydnadsplikt är emellertid komplext. Om läraren anses skyldig att följa en arbetsorder fastän den ligger utanför arbetsgivarens bestämmanderätt, skulle dock eventuella upphovsrättsligt skyddade resultat av sådant arbete tydligt hamna utanför tumregels tillämpningsområde.

[77] I de flesta fall torde dock tvisten uppkomma på en annan nivå, nämligen mellan kursföreståndare – som en slags arbetsgivarföreträdare – och enskilda lärare som kollegor, som beroende snäva budgetramar för den enskilda kursen och snåla avräkningsnycklar för återanvändning har motstående intressen, men då det ut kollegialitetssynvinkel kan vara svårare för upphovspersonen att stå emot ett önskemål om återanvändning som ur rent juridisk synvinkel inte kan tvingas genom.

[78] Se Karnell (1972) s. 36 om betydelsen av att materialet skapats efter beställning från arbetsgivaren.

lärosäten och institutioner om läraren har särskilt avsatt arbetstid (”får timmar”) för sådana insatser, eller om det är något som ska rymmas inom ramen för den förberedelsetid som ges i samband med själva undervisningen eller inom den resterande del av arbetstiden som ska täcka så kallad allmän institutionstjänstgöring och att följa utvecklingen på det egna ämnesområdet. Detta leder oss till frågan huruvida lärosätet genom den lön som betalas till läraren kan anses ha betalt också för undervisningsmaterial som skapas inom ramen för lärarens tjänst.

3.2.3 Beräkning av undervisningstimmar och arbetstid

Att det finns anledning att fästa vikt vid huruvida läraren kan anses ha fått betalt för det upphovsrättsskyddade material som skapats inom ramen för lärarens anställning, hänger samman med att övergång av upphovsrätt till lärosätena i annat fall i realiteten skulle bli vederlagsfritt, något som skulle komma i konflikt med de grundläggande principerna om upphovspersonens ekonomiska ensamrätt såväl som de hänsyn som tumregeln bygger på.[79]

Den arbetsrättsliga utgångspunkten är att man ska få betalt för den tid som man arbetar och att man inte utan särskild ersättning ska arbeta mer än vad som följer av anställningsavtalet.[80] Denna utgångspunkt står ofta långt från den faktiska verkligheten för universitetsläraren, dels därför att många helt frivilligt gärna lägger mer tid på att forska än vad tjänsten inrymmer, dels därför att det som främst räknas för arbetsgivaren är att man genomför sina *undervisningstimmar* – vilka beräknas enligt institutionens egna riktlinjer och oftast oberoende av faktiskt förbrukad tid.[81]

En undervisningstimme motsvarar normalt flera klocktimmar, och är tänkt att ge utrymme för förberedelser och andra undervisningsrelaterade uppgifter, åtminstone när det handlar om salsundervisning. Tilläggas bör att många av de uppgifter som ingår i tjänsten inte finns med i riktlinjerna för beräkning av undervisningstid, fastän det kan handla om

[79] Se förutom AD 2002 nr 87 och Karnell (1972) s. 35, t.ex. Glavå & Hansson (2016) s. 509 ff.

[80] Se Källström & Malmberg (2016) s. 204 f. Ersättningsfrågor är dock inte reglerade i lag i svensk rätt, utan följer normalt av kollektivavtal. Se Glavå & Hansson (2016) s. 602.

[81] Det varierar hur stor andel av tjänsten som utgör forskning, men åtminstone vid juridiska institutioner i Sverige är det inte ovanligt att undervisningsskyldigheten utgör 70–80 % av arbetstiden och att gränsen mellan arbetstid och fritid suddas ut. Se något närmare om verk som skapats utom arbetstid Svensäter (1991) s. 389 f.

undervisningsrelaterade uppgifter. Huruvida den enskilde faktiskt behövt använda all förberedelsetid för att kunna genomföra undervisningen väl eller inte, eller behövt ännu mer, varierar från person till person. Normalt sker dock ingen redovisning av arbetstid utöver den som gottskrivs som undervisningstid, och huruvida det alls finns utrymme inom ramen för redan tilldelad undervisningstid att genomföra andra uppgifter än den aktuella undervisningen kan i normalfallet bara den anställde själv ha någon uppfattning om.

Den lärare som är ny på en kurs får till exempel räkna med att lägga mer tid på att förbereda och genomföra sin undervisning än vad som ingår i tjänsten, medan den erfarna läraren som undervisat på samma kurs i flera år kan gå ”med vinst” varje gång som kursen ges. Att man får samma antal timmar i förberedelsetid oavsett om man är oerfaren eller rutinerad, och oavsett om man genomför samma undervisning år efter år eller om man växlar mellan kurser och ständigt behöver sätta sig in i nya sådana, och kanske till och med behöver skapa nytt material, är ett system som kan leda till suboptimering.[82] Samtidigt anser nog de flesta att verksamheten inte skulle fungera väl om man i stället beräknade arbetstid utifrån principer om faktiskt nedlagd tid för de olika uppgifterna som ska utföras. Snarare skulle ett sådant system sannolikt leda till en ledningsmardröm och närmast oöverstigliga samordningsutmaningar, och även då skulle en påtaglig risk för suboptimering föreligga.[83]

Om läraren får undervisningstimmar för att skapa material, exempelvis seminarie- eller tentamensuppgifter, torde det emellertid ligga inom arbetsgivarens arbetsledningsrätt att ålägga läraren en sådan uppgift på samma sätt som arbetsgivaren kan kräva att läraren håller en bestämd föreläsning eller genomför seminarieundervisning inom ramen för dennes undervisningsskyldighet.[84] Läraren skulle då vara skyldig att göra sitt

[82] Det blir i hög grad upp till den enskilde att själv avgöra hur mycket tid som läggs på att utveckla och förbättra undervisningsmetoder, material och sig själv som lärare – och sådant arbete ger som huvudregel inga timmar. Den som gör minst möjligt av det som inte mäts i inrapporteringen av undervisningstimmar får mer tid till forskning och annat än den som använder ”överbliven” tid från förberedelserna eller mer för att utveckla och förbättra undervisningsmaterial.

[83] Eftersom något yttre incitament till effektivisering av arbetet då skulle saknas, vore risken att onödigt mycket tid läggs på att genomföra arbetsuppgifterna eftersom detta skulle leda till att den totala arbetsbördan reduceras.

[84] Särskild ersättning, om än i form av timmar, som uttryckligen avser framställning av ett bestämt resultat får i sammanhanget anses som en beställning, se Karnell (1972) s. 36.

bästa för att skapa ett – enligt *sin* uppfattning – så bra resultat som möjligt i ett – enligt *sin* uppfattning – lämpligt format.

Det som rent arbetsrättsligt kan krävas är emellertid att läraren gör sitt bästa för att fullgöra sina uppgifter inom de *antal timmar som ryms* i den del av tjänsten som hänför sig till undervisningsskyldigheten.[85] Detsamma gäller när framställningen av material sker efter en särskild beställning. För de flesta lärare torde det dock vara uteslutet att skicka iväg en text eller hålla en föreläsning innan man är nöjd med resultatet, fastän detta kan kräva väsentligt mer tid än den som ges i avräkning. I synnerhet torde detta gälla om det rör material som lätt kan återanvändas, och som riskerar att spridas i större utsträckning än vad läraren tänkt eller kan kontrollera.[86] Här görs en av de tankar som är särpräglad för upphovsrätten och som motiverar det ideella skyddet för upphovspersonen tydlig, nämligen att den sorts verk som skyddas gärna uppfattas som något *personligt* för upphovspersonen.[87]

Fastän arbetsgivaren kan ålägga läraren att på ett visst antal timmar göra sitt bästa för att skapa önskat material, kan läraren *inte* tvingas att lämna ifrån sig arbetet om denne inte blivit nöjd med resultatet efter att ha utnyttjat den tilldelade tiden.[88] På denna punkt skiljer sig salsundervisning tydligt åt från skapandet av material som kan återanvändas. Det är inte tveksamt att läraren kan åläggas genomföra viss undervisning inom den ämnesinriktning som anställningsavtalet anger, och det gäller oavsett om läraren känner sig redo för det aktuella undervisningspasset eller inte. Det kan visserligen vara krävande att behöva stå i en sal och föreläsa om förberedelsetiden inte har räckt till, men en sådan situation är relativt snabbt avklarad och något som i en mycket begränsad utsträckning följer läraren vidare på det sätt som verk framställda i skrift eller som är fångade på ljud eller film kan göra. Fastän det normalt inte framgår vem som skapat seminarie- eller tentamensuppgifter är det internt ofta

[85] Se Glavå & Hansson (2016) s. 506 om arbetstagarens allmänna arbetsrättsliga skyldigheter.

[86] Att det inte innebär åsidosättande av ens skyldigheter enligt anställningsavtalet att lägga ytterligare arbetstid på sådana uppgifter borde vara uppenbart – åtminstone inom den tidsram som står till något så nära fritt förfogande. Mer problematiskt blir det om man skulle behöva använda sin forskningstid, detta eftersom forskning också ingår i arbetsbeskrivningen för de flesta lärare.

[87] Se SOU 1956:25 s. 85.

[88] Det är trots allt upphovspersonen som avgör när ett verk ska offentliggöras, se Bernitz m.fl. (2020) s. 73 ff. jfr NJA 1985 s. 893.

känt vem som skapat vilka uppgifter. Detta innebär att läraren kan ha ett lika skyddsvärt behov av att inte tvingas offentliggöra ett sådant verk förrän han eller hon anser det redo för offentliggörande, som vid skapandet av andra verk.

Visserligen vill lärare normalt göra sitt yttersta för att skapa bra undervisningsmaterial oberoende av om timmarna som utgår täcker den krävda arbetsinsatsen eller inte. Men om beräkningsnyckeln blir alltför snål skulle det, särskilt i kombination med en utövning av arbetsledningsrätten, kunna leda till mycket dåliga resultat. Mot denna bakgrund är det inte överraskande att det gärna blir den enskilde lärarens val om den vill åta sig att skapa material som kan återanvändas, och att sådant skapande i så fall sker närmast oberoende av det antal timmar som utgår för arbetet. Den som har ett långsiktigt perspektiv vet att det extra arbete som utförs vid en tidpunkt, åtminstone till viss del betalar sig senare, till exempel genom att man över flera år har skapat material som kan justeras, omarbetas och återanvändas, och att behovet för förberedelser reduceras när man väl undervisat ett moment några gånger. Den som är villig att lägga ner extra tid då nytt undervisningsmaterial skapas skulle också kunna tänkas ”få tillbaka” tiden senare genom att återanvända materialet till en lägre tidsmässig insats än timmarna som då utgår.[89]

Systemet förutsätter emellertid att de anställda accepterar att risken för såväl kvalitet som tidsåtgång i någon mening ligger på den enskilde. Detta kräver en hög grad av tillit mellan arbetsgivare och arbetstagare, åtminstone om man vill främja innovation och utveckling av nytt material och nya undervisningsmoment.[90] Denna praktiska verklighet innebär också att det, både för lärare och lärosäte, kan vara oklart och nära omöjligt att identifiera om, eller när, läraren kan anses ha ”tjänat in” den tid som gått åt för att skapa visst undervisningsmaterial. Ytterligare kom-

[89] Genom att besluta om egna beräkningsnycklar för återanvändningen av förinspelat material kan man ge tydliga incitament som antingen främjar eller förhindrar att en lärare lägger den tid som behövs för att skapa nytt och för ändamålet lämpligt material, eftersom denne då redan på förhand kommer att veta om man kommer, eller inte kommer, få avräkning för insatsen längre fram.

[90] Just förhållandet mellan upphovsrätten och önskemålen om innovation, ökad värdeskapning och hantering av de stora immateriella tillgångarna inom högskolesektorn har tydligt kommit till uttryck i flera av de utredningar som genomförts på området. Se t.ex. SOU 2020:59 Innovation som drivkraft – från forskning till nytta s. 67 med vidare hänvisning till SOU 1992:7 Kompetensutveckling en nationell strategi och SOU 1998:128 Samverkan mellan universitet och samhället i övrigt.

plicerad blir bedömningen allt eftersom uppdateringar krävs eller andra ändringsbehov uppkommer, vilka i sig leder till ytterligare tidsåtgång. Att identifiera om upphovsrättsskyddat material som skapats inom ramen för en universitetslärares tjänst också rent faktiskt täckts in av den arbetstid som lönen avses utgöra betalning för, skulle alltså förutsätta konkreta, och sannolikt komplexa, bedömningar i det enskilda fallet. I sig utgör detta ett tydligt argument mot att någon förfoganderätt till undervisningsmaterial skapat inom ramen för en tjänst rent allmänt ska övergå till lärosätet med stöd av tumregeln.

3.2.4 Lärosätenas behov av att kunna nyttja undervisningsmaterial skapat av dess lärare

Särskilt på större utbildningar kan det vara nödvändigt med en hel del undervisningsmaterial som ska kunna användas av olika lärare, och gärna över längre tid. När sådant material skapas är det ofta förutsatt från både lärosätets och lärarens sida att det kommer att användas på ett specifikt sätt som också innebär återanvändning, och det är inte ovanligt att upphovspersonens namn inte ens framgår när uppgifterna används.[91] Detsamma gäller vid tentamensuppgifter och i vissa fall även tentamenskommentarer, som på grund av deras karaktär som examinerande moment och därmed underlag för myndighetsutövning kommer att behöva hållas tillgängliga även efter att de använts.[92] Ett av de hänsyn som tumregeln bygger på är arbetsgivarens legitima *behov* av att för sin verksamhet erhålla nyttjanderätt till det upphovsrättsskyddade material som skapats av de anställda i utbyte mot den lön som betalas.[93]

[91] Vilket ju, i någon mening, utgör ett brott med den tydliga upphovsrättsliga huvudregeln om att upphovspersonen enligt 1 kap. 3 § upphovsrättslagen ska namnges. Samtidigt kan man se det så att det i dessa fall finns en sedvänja som innebär att det är i linje med god sed att inte namnge upphovspersonen i materialet som görs tillgängligt för studenterna. De flesta lärare på utbildningar med stora studentkullar skulle nog också kunna se vissa fördelar med att inte bli identifierade.

[92] Denna sorts material får anses utgöra allmän offentlig handling som kan begäras ut av envar, vilket dock inte ger rätt att använda materialet på ett sätt som innebär upphovsrättsintrång.

[93] Se något närmare Svensäter (1991) s. 362 ff. om behovskriteriet. Flera av de riktlinjer som man kan hitta exempel på hos svenska lärosäten idag förespråkar en helt onyanserad och väsentligt mer långtgående övergång av upphovsrätt än vad som torde följa av tumregeln. Tydliga riktlinjer kan visserligen medverka till att det bildas en sedvänja kring vad som gäller, vilken dels kan tänkas utgöra tolkningsunderlag och dels ge innehåll till en

Man kan tänka sig olika sätt på vilka ett nyttjande av undervisningsmaterial skulle kunna vara av betydelse för lärosätet. Den kanske mest långtgående varianten är att lärosätet vid undervisningen av nya studenter vid senare terminer rentav återanvänder material som har skapats av lärare i ett tidigare skede. Ett annat sätt att nyttiggöra sig undervisningsmaterialet handlar snarare om internt bruk för att säkerställa kontinuitet i verksamheten och underlätta för nya lärare. Till exempel kan tidigare föreläsningsbilder vara av stort värde för den lärare som är ny på en kurs och som själv ska skapa nya föreläsningar eftersom bilderna ger en översikt över de ämnen och den tematik som tidigare hanterats inom ramen för den aktuella föreläsningen eller den aktuella kursen. På motsvarande sätt kan förinspelade föreläsningar, seminarieuppgifter eller annat material användas internt, utan att det innebär att materialet används igen *i undervisningen*.

Denna sorts informationsförmedling *kan* i vissa fall tänkas ske utan att det förutsätter någon exemplarframställning utöver den som redan gjorts av den ursprunglige läraren i samband med uppladdning på lärosätets lärplattform vid en tidigare kursomgång, dock förutsätter det ofta att en digital kopia skapas vid nedladdning och eventuell vidarebefordran till aktuell personal. Gemensamt är att det handlar om användning för enskilt bruk, om än inom ramen för verksamheten. Upphovsrättsligt gäller emellertid som huvudregel samma utgångspunkt för exemplarframställning så länge användningen sker som del av en professionell verksamhet och inte för privat bruk.[94] Huruvida det handlar om rent intern användning eller dessutom användning i undervisningen kan dock ha betydelse när man ska avgöra frågan om övergång av nyttjanderätt med stöd av tumregeln. Fortsatt intern användning torde nämligen vara väsentligt mer vanligt förekommande, och för upphovspersonen mindre kränkande, än återanvändning i undervisningssituationer.

I normalfallet, åtminstone så länge som upphovspersonen är fortsatt anställd vid lärosätet, torde det vara sällsynt förekommande att läraren skulle vilja förhindra att material som denne har skapat används som

eventuell tumregel på området. Det är emellertid inte nog att riktlinjerna finns och står där utan att ifrågasättas, utan man måste i så fall också visa att de rent faktiskt efterlevs över tid. Än så länge är det litet som tyder på att så är fallet för de riktlinjer som gäller vid svenska lärosäten idag.

[94] Se närmare Olsson & Rosén (2018) avsnittet om 2 kap. 12 § första stycket upphovsrättslagen om ändringarna i bestämmelsen till följd av införlivandet av Infosocdirektivet (2001/29/EG) och kopplingen till reglerna om avtalslicenser.

inspiration för kollegor eller till annat internt bruk. Gäller det enklare föreläsningsbilder, seminarie- eller tentamensuppgifter som redan ”förbrukats” bör samma sak gälla även efter att läraren har avslutat sin anställning, och för seminarie- och tentamensuppgifter sannolikt oberoende av om materialet används vidare i verksamheten riktad till studenterna eller endast internt.

I dessa fall kan man tänka sig att lärosätet i viss mån äger rätten att nyttja materialet eftersom det lämnats över till lärosätet under förutsättning att det används på ett bestämt sätt, åtminstone om läraren har fått timmar för skapandet. För insatser som ersätts utifrån fasta tariffer för beräkning av undervisningstid, kan det presumeras att lärosätet erhåller viss nyttjanderätt till materialet som skapats, som i vissa fall också innefattar återanvändning i undervisningssituationen. Om man vill kalla detta för en följd av en tumregel på området, eller som uttryck för ett underförstått avtal, blir främst en semantisk fråga eftersom avtalet i ett sådant fall – i brist på konkreta hållpunkter om annat – får anses innebära ett i tid obegränsat samtycke till nyttjande på det sätt som var vanligt i verksamheten vid avtalstidpunkten.[95]

Detta kan dock sägas utgöra en utgångspunkt endast om det handlar om en mer konkret *beställning* från arbetsgivarens sida vilken *ryms* inom arbetsledningsrätten,[96] vilket alltså förutsätter att lärarens insats åtminstone rimligen ryms inom dennes arbetstid. Om situationen präglas av att läraren felaktigt uppfattat sig vara förpliktad att skapa material utan att det gottskrivs som undervisningstid, är risken att läraren implicit åläggs att jobba mer än vad som krävs enligt arbetsskyldigheten utan ersättning för det. I dessa fall skulle en övergång av upphovsrätt till följd av anställningsförhållandet inte heller kunna motiveras med de hänsyn som utgör skäl för tumregeln.

Också om det gäller mer avancerade undervisningsupplägg som skapats på lärarens eget initiativ, eller om det rör material avsett för den aktuella lärarens eget bruk, är det nog ofta naturligt att tänka sig att läraren har ett intresse av att begränsa lärosätets användning av materialet. Exempel i sammanhanget är digitala spel och liknande uppgifter eller förinspelat material där lärarens egen röst och eventuellt bild finns med. Detsamma gäller föreläsningar och stöddokument, oavsett format och

[95] Se liknande Karnell (1972) s. 36 om förhållandet mellan tumregeln och beställningsfallen.

[96] Se Karnell (1972) s. 35–36.

framförande, som oftast också utgör verk som är avsett för just den egna personliga användningen. I dessa fall får utgångspunkten anses vara att den upphovsrätt som eventuellt övergår till arbetsgivaren som *mest* kan omfatta rätt att nyttja materialet internt i den mån det behövs för att säkerställa verksamhetens kontinuitet och tjäna som inspiration för övriga anställda. Någon mer omfattande nyttjanderätt, till exempel för fortsatt användning i undervisningen, ryms som jag ser det inte av någon tumregel på området utan skulle kräva att det finns konkreta hållpunkter för att en sådan övergång varit förutsatt mellan de båda parterna.[97]

Gränsfall kan tänkas finnas, till exempel om läraren erhållit viss tilldelning av timmar för att skapa visst material som också förutsätter en insats vilken kräver arbete utöver den tilldelningen eller vad ordinarie arbetstid omfattar, men det leder oss i så fall över till bedömningar om underförstådda avtal i det enskilda fallet och inte övergång av upphovsrätt på grund av anställningsförhållandet som sådant. På samma sätt som när en tentamensuppgift skapas och används i utbyte mot att läraren erhåller ett på förhand fastställt antal timmar, torde det emellertid ligga nära till hands att tolka situationen så att samtycke till lärosätets nyttjande getts när läraren har åtagit sig och också fullgjort åtagandet. Annorlunda vore det om arbetet genomförts på eget initiativ, där lärosätets eventuella bidrag varit externa medel som använts exempelvis för att täcka lärarens kostnader för att köpa in nödvändig utrustning.

3.2.5 Sammanfattning

Det kan i *mycket begränsad* omfattning uppställas en presumtion för att lärosätet ska erhålla någon nyttjanderätt till det undervisningsmaterial läraren skapat inom ramen för sin tjänst. Kombinationen av akademisk frihet och valfrihet avseende hur den enskilde läraren vill utforma det undervisningsmaterial han eller hon skapar, såväl som individuella skillnader i tidsåtgång och det system för beräkning av arbetstid som är vanligt i branschen, gör att omständigheterna som i andra fall kan motivera övergång av upphovsrätt till följd av anställningsförhållandet, inte gör sig gällande.

I vissa konkreta fall kan man dock identifiera att förutsättningarna för skapandet över lag sannolikt varit sådana att lärosätet kan erhålla rätt

[97] Se också Karnell (1972) s. 38 situation d om att läraren kan motsätta sig återanvändning av material tillkommit för bruk i egen undervisning.

att nyttja materialet även framgent, trots att läraren inte själv samtycker. Dels kan detta anses gälla *internt* bruk av material som tidigare använts i undervisningssituationen, dels återanvändning av material som läraren erhållit undervisningstimmar för att skapa efter konkret förfrågan. Sistnämnda typ av material kan då också, i den utsträckning som det vid tidpunkten för skapandet varit synbart för läraren att återanvändning var normalt i verksamheten, återanvändas i undervisningen även fortsättningsvis.[98]

3.3 Ett upphovsrättsligt lärarundantag för universitetsläraren?

Mot bakgrund av den räckvidd som en tumregel på området enligt min uppfattning kan anses ha framstår behovet av ett upphovsrättsligt lärarundantag som tämligen begränsat. Den övergång av upphovsrätt som förefaller kunna ske med stöd av en tumregel på området motsvarar i grova drag också de begränsningar som sannolikt oavsett måste uppställas i ett upphovsrättsligt lärarundantag.[99] En tvärsäker och absolut syn om att varje övergång av upphovsrätt till lärosätena kräver samtycke,[100] kan jag inte se stöd för varken i rättskällorna eller de intressen som förtjänar att beaktas.

Att man inte undviker behovet av att göra en konkret bedömning av huruvida någon upphovsrätt gått över till följd av universitetslärarens anställningsförhållande genom att åberopa lärarundantaget, innebär att det i rättsteknisk mening spelar mindre roll under vilken rubrik man gör bedömningen. Emellertid framstår det för mig som rent strukturellt lämpligt att först göra bedömningen av om någon upphovsrätt alls kan övergå till lärosätet till följd av anställningsförhållandet, för att därefter bedöma huruvida det kan göras undantag från detta med stöd i lärarundantaget. Mot bakgrund av framställningen ovan, skulle en sådan prövningsordning leda till ett begränsat behov av lärarundantaget, eftersom det sannolikt inte skulle ge universitetsläraren något utökat skydd än vad som redan

[98] De flesta riktlinjerna som beslutats vid olika svenska lärosäten skiljer inte på olika sorters användning och ger därför uttryck för en tumregel med väsentligt större räckvidd än det finns fog för.

[99] Sandgren (2003) s. 39–40.

[100] En så absolut syn framgår främst av Wolk (2011) i hennes rapport för SULF.

följer av tumregeln. Åtminstone gäller det såsom rättsläget och universitetslärarens arbetssituation ser ut *idag*.

Arbetsgivaren kan givetvis, genom att strukturera om sin verksamhet över tid, tänkas påverka verksamhetens karaktär på så sätt att också tumregelns räckvidd kommer att förändras till nackdel för lärarens ensamrätt till sina verk.[101] För denna tänkbara framtida förändring skulle det dock, som jag ser det, krävas mycket drastiska ändringar av verksamheten för att någon egentligt utökad möjlighet till övergång av upphovsrätt genom universitetslärarens anställning. Sannolikt skulle kombinationen av begränsningar i rätten att instruera läraren i utövandet av arbetet, tillsammans med sättet att beräkna arbetstid, också innebära att en ändring i hur man genom verksamheten vill nyttiggöra sig sina anställdas verk inte nödvändigtvis i sig skulle leda till ett väsentligt förändrat resultat vid tilllämpning av tumregeln.

Skulle lärarens arbetssituation ändras på så sätt att de arbetsrättsliga argumenten mot övergång av upphovsrätt till följd av anställningsförhållandet bleknar, skulle man dock kunna tänka sig att tumregeln kan ge stöd för en ökad övergång av upphovsrätt. Det kan därför inte uteslutas att ett upphovsrättsligt lärarundantag skulle kunna få självständig betydelse. Samtidigt skulle det fortfarande, om nu inte verksamheten fullkomligt ändrar karaktär och betydande lagändringar genomförs, handla om en möjligt utvidgad rätt att nyttja just *undervisningsmaterial*, men inte forskning eller andra publikationer.[102]

Att det oavsett finns ett upphovsrättsligt lärarundantag som – oberoende av tumregelns räckvidd – innebär att någon övergång av upphovsrätt från läraren till lärosätet inte sker för forskning eller andra publikationer, torde dock vara uppenbart i dagsläget. Fastän det saknas auktoritativa rättskällor som lärarundantagets existens eller omfattning kan härledas ur, har undantaget hanterats som en säker sedvanerättslig regel i

[101] Se Rosenmeier (2016) s. 701.

[102] Interna riktlinjer kan såklart framöver tänkas spela roll som del i en framväxande sedvänja med betydelse för den sorts material som ligger inom ramen för vad som potentiellt kan omfattas av den sedvanerättsliga tumregeln. För material som oavsett hamnar utanför en tumregel, som till exempel forskning och andra publikationer, dyker emellertid relationen till 2 kap. 16 § RF och 1 kap. 6 § 2 st. 3 högskolelagen upp, vilka torde innebära ett behov av stöd i lag eller enskilt avtal om upphovsrätt till vetenskapliga texter eller läromedel skulle övergå till arbetsgivaren. Se närmare Olsson & Rosén (2018) avsnittet om Upphovsrättssystemets kulturella och ekonomiska roll m.m. om regleringen i 2 kap. 16 § RF.

ett stort antal utredningar och i doktrinen under mycket lång tid. Framställningarna innehåller visserligen var för sig i begränsad utsträckning någon djupare analys av undantagets innehåll eller räckvidd och flera av dem upprepar bara påståenden som hämtats från andra ställen, vilket tydligt begränsar deras vikt som självständiga rättskällor. Trots den kritik som kan riktas mot det rättskällestöd som (inte) anförts till stöd för ett upphovsrättsligt lärarundantag, framstår det ändock inte som tveksamt att det existerar ett sådant undantag på sedvanerättslig grund när det gäller forskning och läromedelsproduktion. När det gäller annat material som skapas av läraren inom ramen för hans eller hennes tjänst, förefaller det däremot för mig som tydligt att frågan om övergång av upphovsrätt till följd av tumregeln och ett eventuellt lärarundantag är två sidor av samma mynt, som torde leda till samma resultat.

Silvia A. Carretta

Liability for Copyright Infringement and Algorithmic Content Moderation: A Matter of Proportion

1 Intro

Every day, a staggering amount of new online content is generated. Only a small amount of that online content is subjected to some form of editorial review before it is posted. Most of that online content is user-generated, from tweets to all kinds of different media uploads, which are often posted without any form of scrutiny from another human being checking the lawfulness of that content. Just to comprehend the sheer size of it, one set of estimates found that every minute, Facebook users upload 147.000 photos; users upload 500 hours of video on YouTube; Reddit sees 479.452 people engage with its content; 456.000 tweets are posted on Twitter.[1] In order to prevent the Internet from being flooded with unreliable misinformation, child pornography, copyright-infringing material and other harmful or unlawful content, an increasing amount of regulation is generated to cleanse the Internet. However, given the enormous amount of online content generated, the task of editorial overview over all of this content goes beyond human capacity. One would need another world population to monitor, review and censor the online content produced by our current world population. Where is the Heracles that can clean the Augean stables of our present-day Internet? And who should be held liable for the uploading of illegal content? In this contribution, I will explore this question in relation to the content

[1] For an infographic on these impressive data, see: Domo, Data never sleeps 8.0 (2020). <domo.com/learn/infographic/data-never-sleeps-8> accessed 7 August 2021.

moderation of copyright-infringing material and the controversial Article 17 of the Directive (EU) 2019/790 on Copyright in the Digital Single Market (DSM Directive)[2] which establishes a platform liability that, in practice, seems difficult to avoid without the platform taking recourse to automated upload-filters.

The concept of copyright protection is to protect creativity and give copyright holders the power to control reproduction and communication of their works to the public. The enormous amount of user-generated content, as mentioned above, can infringe on these rights when it includes copyright protected content (pictures, text, music, videos, etc.) that are shared in a way that makes them available for viewing, downloading or online distribution.[3]

Online content sharing has become the subject of extensive regulation, in order to protect copyright and limit the widespread issue of piracy in digital space.[4] As argued by Gorwa, "*copyright has historically been one of the first, if not the first, domain where strong economic interests demanded technologies to match and classify online content*"[5]. In copyright law, a distinction is made between primary liability for individual copyright infringers, that is, the users that upload copyright protected material, and secondary liability for third-party intermediaries that facilitate the users in their copyright infringements, for example, platforms like Facebook or Twitter. The EU legislative framework already contains instruments that establish the primary and secondary liability for copyright infringements. In particular, most jurisdictions have provisions whereby third parties can be held liable for contributing to copyright infringement by their

[2] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

[3] Brian Fitzgerald et al. Search Engine Liability for Copyright Infringement, in Amanda Spink, Michael Zimmer (eds), *Web Search: Multidisciplinary Perspectives* (Information Science and Knowledge Management, vol. 14 Springer 2008), 104.

[4] As per a EUIPO study from 2018, piracy remains a significant problem within the EU, where the average Internet user accesses pirated content 9.7 times per month in 2018. In fact, pirated video-material gets over 230 billion views a year, and more than 80% of global online piracy can be attributed to illegal streaming services. See: <dataprot.net/statistics/piracy-statistics/> and EUIPO, Online copyright infringement in the European Union music, films and tv (2017–2018). Trends and drivers (2019). <euipo.europa.eu/online_copyright_infringement_in_eu_en.pdf> accessed 7 August 2021.

[5] Robert Gorwa et al. *Algorithmic content moderation: Technical and political challenges in the automation of platform governance* (Big Data & Society Vol. 7 2020).

users. The rationale behind such provisions is that these third parties are often in a better position to discourage infringement, by implementing mechanisms to monitor infringers' activity or, at least, making it more difficult to share copyright protected works.[6]

In particular, the often mentioned and debated Article 17 DSM Directive contains an obligation for certain type of intermediaries – i.e. Online Content Sharing Service Providers[7] – to ensure that copyright infringing uploads made by their users on their platforms are prevented and/or removed through content-filtering procedures, which are based either on information sent by the rights holders or on automatic preventive filtering. If these Intermediaries are not proactive in moderating content and removing infringing uploads, they could be directly liable for copyright infringement.

The main problem for these Intermediaries in complying with the law is the aforementioned immense number of uploads generated daily by their users. Furthermore, things are complicated by the fact that content may be created in one country and viewed in another, thus requiring that Intermediaries create and enforce different legal requirements and cultural policies for each country.

Such requirements for content moderation can only be manageable, thanks to automated filtering technologies based on artificial intelligence.[8] Artificial Intelligence (AI) can play a fundamental role in fighting online copyright infringements, thanks to its 'predict and prevent' approach: algorithms have the ability to rapidly analyse huge amounts

[6] On the economics of intermediaries' liability for copyright infringement, see: Douglas Lichtman and William Landes, *Indirect Liability for Copyright Infringement: An Economic Perspective* (Harvard Journal of Law & Technology Vol. 6 2003).

[7] Although there is a plethora of online intermediaries, each different definition and functions (e.g. Internet service providers, search engines, social media platforms, web hosting providers, etc.), this paper focuses only on Online Content Sharing Service Providers. These Intermediaries are not a new category of online providers in a technological sense. They are instead a new legal category regulated by a body of provisions from the E-Commerce Directive, the InfoSoc Directive, the Enforcement Directive and the DSM Directive. For the sake of readability, they will be referred to by the general term of 'Intermediaries'.

[8] For a detailed overview of the various filtering technologies, see the study requested by the JURI Committee to Giovanni Sartor, Andrea Loreggia: Policy Department for Citizens' Rights and Constitutional Affairs, *The Impact of algorithms for online content. "Upload Filters"* (2020), 35 <europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2020)657101> accessed 7 August 2021.

of data, identify patterns and proactively make predictions to evaluate if a work contains infringing content. When an algorithm identifies a content as potential copyright infringement, it can automatically proceed with an algorithmic assessment to decide on the best actions to take: these actions are preventive means to either filter the content or to remove the infringing works. The expanding liability of Intermediaries for copyright infringing content posted by their users is pushing the latter to foster the development of AI algorithms (alongside human reviewers) to optimise a speedy detection and removal system.[9]

Nevertheless, AI is not a panacea for all online infringements. The use of AI algorithms by Intermediaries to automatically moderate online content in order to limit their liability has been criticised due to the serious danger of AI likely leading to over-blocking of lawful content, as a collateral effect to automated decision-making. In fact, AI is still incapable of properly interpreting context-related uses, and distinguishing between lawful and unlawful uses, in particular, for cases that might fall under one of the exceptions or limitations provided for by national copyright legislation (such as parody or criticism). Over-blocking based on automated upload filters could thus result in potential limitations to and infringements on fundamental rights (for instance, the right to freedom of expression and of the arts) and basic principles of EU law (such as proportionality and legal certainty).

After having briefly presented the topic of discussion in this introductory Section 1, I will proceed to present in Section 2 the characteristics of AI-based mechanisms for automated algorithmic content moderation and then to introduce technical limitations of AI when used by Intermediaries for the private governance of their platforms. Subsequently, in Section 3, I will illustrate the EU legal framework for Intermediaries' liability, focusing in particular on the application of the aforementioned Article 17 DSM Directive, its safe harbour exceptions – which limit Intermediaries' liability – and the mandatory exceptions and safeguards created to strengthen the rights of copyright holders. Furthermore, I will present a critical analysis of how privately operated algorithms for content moderation might fail to appropriately balance the protection of copyright and fundamental rights, due to inherent limits and flaws of the technology. At last, I will draw conclusions in Section 4, together with

[9] Kirsten Gollatz et al., *The turn to artificial intelligence in governing communication online* (SocArXiv, 2018). <osf.io/preprints/socarxiv/vwpcz> accessed 7 August 2021.

speculations over further development needed for AI-based mechanisms to obtain a suitable balance between copyright protection and legal certainty on Intermediaries' liability.

2 Turning to AI for algorithmic content moderation at scale

As governments, advertisers, and users' pressure on major Intermediaries is growing, both companies and legislators are searching for technical solutions to the difficult puzzle of Intermediaries' governance and online content moderation against copyright infringement. Intermediaries have to handle an enormous volume of data due to the "*stratospheric*" quantity, velocity, and variety of content consumed online,[10] which makes it impossible for them to only rely on prompting human review.

In recent years, AI has been deployed by Intermediaries to reduce the reliance on users to flag content for review, to automatically being able to remove allegedly infringing content, or even filter it out before it is uploaded.[11] As accurately described by Gillespie, "*this link between platforms, moderation, and AI is quickly becoming self-fulfilling: platforms have reached a scale where only AI solutions seem viable; AI solutions allow platforms to grow further.*"[12]

2.1 The promise of AI

AI plays an important role in shaping online content moderation and in helping Intermediaries enforce content moderation with legal certainty, thus determining and proving more clearly where their liability originates or ends. In fact, to stay on the safe side, it might be easier for Intermediaries to hold a 'block-first, verify-later' approach, algorithmically blocking all content that could, even remotely, be infringing copyright. Nevertheless, this would lead to a serious risk of over-blocking lawful contents (as presented in section 3.1 below), which goes against the pur-

[10] Tarleton Gillespie, *Custodians of the Internet: intermediaries, content moderation, and the hidden decisions that shape social media* (Yale University Press 2018).

[11] Op. cit. Kirsten Gollatz (2018).

[12] Tarleton Gillespie, *Content moderation, AI, and the question of scale* (Big Data & Society Vol. 7 2020) 2.

pose of balancing copyright of the rights holders with the rights of users to consume and share lawful content online, without unduly restricting freedoms as collateral effect. This is where algorithmic content moderation and AI come into play.

Algorithmic content moderation can be defined as a governance mechanism enforced by Intermediaries, which use classification of user generated content to implement appropriate choices on how members of a community engage with each other and how content is shared, exploited or removed (e.g. through governance outcomes of removal, geo-blocking, account takedown)[13]. It requires the collection of massive amounts of data from uploaded content and the application of data analytics techniques to identify patterns and make predictions on the best actions to take, to achieve the given governance goals. In the case of online copyright, these goals are to proactively detect, or automatically evaluate whether it contains infringing content, then proceeding with preventive filtering or removal.[14]

Digital and computational methods can be combined usefully with statistical and rich qualitative methods to obtain successful large-scale algorithmic content moderation. These methods span from simpler technical approaches, including keyword filtering (i.e. scanning of a text to identify blacklisted words or phrases stored in a database) and hash matching (i.e. generation of a unique digital fingerprint for previously detected harmful images and videos, to which every new upload is compared to verify its harmfulness),[15] to more sophisticated machine learning-based systems, such as natural language processing (i.e. field of study aiming to enable algorithms to comprehend texts in a more extended

[13] Robert Gorwa et al. *Algorithmic content moderation: Technical and political challenges in the automation of platform governance* (Big Data & Society Vol. 7 2020).

[14] The concepts of proactive detection and automated evaluation are mentioned in a variety of policies and legislations: e.g. "monitoring obligations" (article 15 Directive 2000/31/EC), "notice and stay-down", "upload filtering" (European Digital Rights, Copyright directive: Upload filters strike back. Protecting Digital Freedom (2019), "automatic detection and removal of content" (Conclusions. EUCO 8/17, par. 2, European Council meeting (22 and 23 July 2017). See Emma J. Llansò, *No amount of "AI" in content moderation will solve filtering's prior restraint problem* (Big Data & Society Vol. 7 2020).

[15] See e.g. Microsoft's PhotoDNA tool. <microsoft.com/en-us/photodna> accessed 7 August 2021.

way, closer to the way humans understand text and its context),[16] and optical character recognition (i.e. identification of text in an image and conversion of it into machine-readable format).

Moreover, AI approaches to image and video analysis can be used to detect the presence of pre-identified objects, scenes or elements, such as symbols or logos (i.e. object recognition is the identification of specific pre-defined object classes within an image), for semantic segmentation (i.e. detection and identification of harmful objects and their location by pixels analysis) and scene understanding (i.e. identification of scenes within images, by comparing their dimensional representation to other objects in the image). Other deep-learning methods enable techniques for audio channel separation (i.e. separation of audio sources for deeper analysis) and hash-matching (i.e. identification of audio by comparison to previously categorised audio tracks within a database)[17]. The use of these AI-based technologies allows us to limit the circumvention of filtering by slight alteration of the content in video, images and text (e.g. cropping an image, adding a filter, modifying the lighting conditions or resolutions, rotating/skewing of an element, or modifying the caption could defeat the filter's ability to identify an infringing content).

AI-based algorithms have been deployed in a variety of contexts to protect intellectual property rights. The first example to mention is the 'BookID' system used by Scribd, the subscription-based digital library of e-books and audiobooks. It is described as a system that "*algorithmically analyses computer-readable text for semantic data (such as word counts, letter frequency, phrase comparisons and so on) that it then encodes into a digital "fingerprint". It scans every document uploaded to Scribd and removes those that have the same, or a substantially similar, fingerprint. BookID's approach*

[16] For instance, Google & Jigsaw's Perspective API is an open-source toolkit that allows Intermediaries and users alike to use its machine learning models to evaluate the "toxicity" of a post or comment. <perspectiveapi.com/>accessed 7 August 2021.

[17] For a broader analysis, see: Cambridge Consultants for UK OfCom, Use of AI in Content Moderation, (2019), <ofcom.org.uk/research-and-data/internet-and-on-demand-research/online-content-moderation> accessed 7 August 2021; Emma Llansò et al., Artificial intelligence, content moderation, and freedom of expression, In: *Transatlantic Working Group on Content Moderation Online and Freedom of Expression* (IViR 2019) <ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf> accessed 7 August 2021.

reduces misidentifications and enables the detection of infringing works even if they have been altered to some degree."[18]

Similarly, Amazon's 'Project Zero', powered by a machine learning algorithm, continuously scans product listing updates to proactively remove suspected counterfeits, based on logos, trademarks, and key data provided by its partnering brands.[19]

A final example: YouTube has experimented since 2006 with a voluntary, automated content monitoring system called 'ContentID', which is formally and procedurally independent from its notice-and-takedown-process (legally necessary to satisfy its obligations under the safe harbour regimes and limit its liability for copyright infringing content uploaded by its users). This algorithm operates on a 'predict and prevent' approach, using a digital fingerprint system: it detects the matching between a newly uploaded video and a protected work, going as far as to monitoring live chats and video meta-data to predict whether an audio or video is a copyright-infringing live stream of sports games. Once a match has been found, rights holders are notified, and they have the ability to either block or take down the content[20]; a third option is to receive a portion of the advertising revenue from the uploaded content.

2.2 Technical limitations and conflicts of private governance of platforms

In an interesting analysis, Elkin-Koren presents how "*overall, content moderation by AI reflects the rise of unchecked private power, which may escape traditional checks and balances intended to ensure that power is exercised in the interest of society at large*"[21]. As automated, privatised, algorith-

[18] Scribd, About the BookID™ Copyright Protection System, 2021 <support.scribd.com/hc/en-us/articles/360037497152-About-the-BookID-Copyright-Protection-System> accessed 7 August 2021.

[19] Amazon, Project Zero leverages the combined strengths of Amazon and brands to drive counterfeits to zero, 2021 <brandservices.amazon.se/projectzero> accessed 7 August 2021.

[20] According to YouTube "*over 98% of copyright issues are handled through Content ID, rather than the notice-and-takedown process. […] as it automatically identified the work and applied the copyright owner's preferred action*". See Google, How Google Fights Piracy, 2018 <storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Piracy_2018.pdf> accessed 7 August 2021.

[21] Niva Elkin-Koren, *Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence* (Big Data & Society Vol. 7 2020), p. 2.

mic content moderation systems are more frequently used by Intermediaries for online governance, scholars have raised the concern that these algorithms might take over decision-making power normally assigned to courts and administrative agents.[22] In fact, as described by Heldt in a recent work, algorithmic content moderation systems are endowed with censorial power, bypassing traditional checks and balances secured by the law.[23]

Anticipating what will be discussed more in depth in the following section, it is noteworthy to highlight how these AI-based moderation systems are controversial because of the risk of unduly restricting freedom of expression and of the arts, which are bestowed over users to access, experience and share creative content online (e.g. through scientific publications, cultural assets and news reports)[24].

Consequently, another issue arises from the potential negative outcomes of using algorithms for copyright enforcement, due to prioritization of efficiency over accuracy,[25] which might lead to potential misidentification and over-blocking. As seen further on, algorithmic content moderation faces extensive challenges when context needs to be considered to interpret the meaning of different formats (such as text, images, video or audio). In fact, these heavily automated systems lack contextual sensitivity, having difficulty in identifying subtlety, sarcasm and subcultural meaning,[26] as well as in detecting context and exceptions or limitations provided by the law.

First, limitations arise from the opaqueness of AI algorithms, which makes it harder to ensure that users' rights are adequately protected. There are multiple sources of opacity: to begin with, algorithms and data are often protected as trade secrets, preventing public access in order to

[22] Adam Bridy, Copyright's digital deputies: DMCA-plus enforcement by internet intermediaries. In: John Rothchild (eds.) *Research Handbook on Electronic Commerce Law* (2016), 185–208.

[23] Amélie Pia Heldt, *Upload-filters: Bypassing classical concepts of censorship* (JIPITEC 10 (1) 2019).

[24] Op. cit. Emma Llansò (2019).

[25] Joanne E. Gray, *Google Rules: The History and Future of Copyright under the Influence of Google* (Oxford University Press 2020).

[26] Natasha Duarte et al., *Mixed messages? The limits of automated social media content analysis*, Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR 81:106–106, 2018) <cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf> accessed 7 August 2021.

determine why one piece of content, rather than another, has been subjected to removal.

Secondly, automated decision-making frequently occurs by means of so-called 'black box' algorithms whose predictions are often difficult to interpret – even for the data scientists who designed the system. Moreover, there are additional concerns about transparency, accountability and protection of fundamental rights since the data and training models are often kept confidential by Intermediaries who seek to avoid public scrutiny. In fact, machine learning algorithms are only as good as the datasets they are trained on: if training data does not include a representative number of examples of different languages and different groups or minorities, there will be significant risks of bias and erroneous classifications of underrepresented groups.[27]

Finally, the accuracy of these algorithms is oftentimes embellished: while some Intermediaries indeed use AI for preventive content moderation, most only use a sophisticated version of hash/pattern matching, which could hardly be included under the definition of AI, except under the broadest possible one[28].

3 Principles of secondary liability of online Intermediaries

As governments and users alike continue to strengthen their pressure over Intermediaries to take a more active role in moderating online content, it becomes increasingly important to deploy proper mechanisms to hold Intermediaries to account for copyright infringements originating on their platforms. Throughout the years, various EU and national copyright legislations[29] introduced provisions to encourage Intermediaries to

[27] Op. cit. Tarleton Gillespie (2018).

[28] Researcher Julian Togelius addresses this question well in his blog post: "*There is no such thing as an artificial intelligence. AI is a collection of methods and ideas for building software that can do some of the things that humans can do with their brains. Researchers and developers develop new AI methods (and use existing AI methods) to build software (and sometimes also hardware) that can do something impressive, such as playing a game or drawing pictures of cats*". See Julian Togelius, Some advice for journalists writing about artificial intelligence (2019) <togelius.blogspot.com/2017/07/some-advice-for-journalists-writing.html> accessed 7 August 2021.

[29] In the EU, safe harbour provisions were initially introduced with the 'E-Commerce' Directive 2000/31/EC.

develop automated mechanisms to moderate online content in exchange for exemption from liability for content posted by their users. Recently, this Intermediaries' liability regime has been revised by Article 17 of the DSM Directive. This provision is indeed one of its most controversial provisions, and its new liability regime can be summarised as follows.

First and foremost, what is special about Article 17 is not the characterisation that certain Intermediaries perform acts restricted by copyright,[30] rather how it treats the liability of these Intermediaries.[31] Article 17(1)

[30] This is already provided for by Article 3 of the 'InfoSoc' Directive 2000/29/EC, and CJEU case law.

[31] Article 17(1) to (4) DSM Directive on the use of protected content by online content-sharing service providers states that:
"*1. Member States shall provide that an online content-sharing service provider performs an act of communication to the public or an act of making available to the public for the purposes of this Directive when it gives the public access to copyright-protected works or other protected subject matter uploaded by its users.*

An online content-sharing service provider shall therefore obtain an authorisation from the rightholders referred to in Article 3(1) and (2) of Directive 2001/29/EC, for instance by concluding a licensing agreement, in order to communicate to the public or make available to the public works or other subject matter.

2. Member States shall provide that, where an online content-sharing service provider obtains an authorisation, for instance by concluding a licensing agreement, that authorisation shall also cover acts carried out by users of the services falling within the scope of Article 3 of Directive 2001/29/EC when they are not acting on a commercial basis or where their activity does not generate significant revenues.

3. When an online content-sharing service provider performs an act of communication to the public or an act of making available to the public under the conditions laid down in this Directive, the limitation of liability established in Article 14(1) of Directive 2000/31/EC shall not apply to the situations covered by this Article.

4. If no authorisation is granted, online content-sharing service providers shall be liable for unauthorised acts of communication to the public, including making available to the public, of copyright-protected works and other subject matter, unless the service providers demonstrate that they have:

(a) made best efforts to obtain an authorisation, and

(b) made, in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works and other subject matter for which the rightholders have provided the service providers with the relevant and necessary information; and in any event

(c) acted expeditiously, upon receiving a sufficiently substantiated notice from the rightholders, to disable access to, or to remove from their websites, the notified works or other subject matter, and made best efforts to prevent their future uploads in accordance with point (b)."

establishes primary liability for acts of communication, or making available to the public jointly committed by the Intermediaries and its users, which morphs into secondary liability under par. (4) where Intermediaries are liable for infringing content uploaded by their users when failing to obtain the necessary authorisation from rights holders. However, Article 17(4) provides three conditions for Intermediaries to escape liability, by demonstrating to have: i) undertaken 'best efforts' to obtain authorisation; or ii) made "*in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works and other subject matter for which the rights holders have provided the service providers with the relevant and necessary information*" (par. (4)b); iii) acted expeditiously, subsequent to notice from rights holders, to take down infringing content and made best efforts to prevent its future upload[32].

Having initially favoured licensing agreements and preventive authorisations to limit liability, par. (7) of Article 17[33] introduces new mandatory exceptions and limitations applicable to user uploads so that the latter can safely share content online by relying on a general exception for quotation, criticism, review or use for the purpose of caricature, parody or pastiche.[34] Furthermore, a clarification in par. (8) specifies that Article 17 does not entail general monitoring obligations.

[32] This condition seems to introduce a notice-and-takedown mechanism, similar to that of Article 14 E-Commerce Directive and a notice-and-stay-down (or re-upload filtering) obligation for Intermediaries.

[33] Article 17(7) DSM Directive states: "*The cooperation between online content-sharing service providers and rightholders shall not result in the prevention of the availability of works or other subject matter uploaded by users, which do not infringe copyright and related rights, including where such works or other subject matter are covered by an exception or limitation. Member States shall ensure that users in each Member State are able to rely on any of the following existing exceptions or limitations when uploading and making available content generated by users on online content-sharing services: (a) quotation, criticism, review; (b) use for the purpose of caricature, parody or pastiche*".

[34] These new mandatory exceptions and limitations operate alongside the one provided for by Article 5(3) of 'InfoSoc' Directive. In situations of conflict (i.e. an exception is explicitly mentioned in Article 17(7) but unavailable at the national level ex InfoSoc Directive), the former creates an obligation under EU law to transpose under national legislation these exceptions and limitations.

Lastly, par. (9)[35] introduces three safeguards to protect users and to minimise the risks of broad filtering and over-blocking.[36] First, any request by rights holders for the removal of specific content must be justified; second, it requires Member States (while transposing the Directive) to ensure that Intermediaries put in place "*effective and expeditious complaint and redress mechanisms*" which users can avail themselves of in case of disputes over contentious "*decisions to disable access to or remove uploaded content*" (which should be subject to human review); third, Member States should create out-of-court dispute settlement mechanisms, which are independent of the judicial redress. This is in order to guarantee that Intermediaries optimise algorithmic mechanisms for the uniform protection of fundamental rights and freedoms across the EU.[37]

After having presented the novelties of Article 17, I now introduce two technical issues arising from practical application of this provision: i) the risk of over-blocking due to automated preventive filtering and ii) the

[35] As written in Article 17(9) DSM Directive: "*Member States shall provide that online content-sharing service providers put in place an effective and expeditious complaint and redress mechanism that is available to users of their services in the event of disputes over the disabling of access to, or the removal of, works or other subject matter uploaded by them.*

Where rightholders request to have access to their specific works or other subject matter disabled or to have those works or other subject matter removed, they shall duly justify the reasons for their requests. Complaints submitted under the mechanism provided for in the first subparagraph shall be processed without undue delay, and decisions to disable access to or remove uploaded content shall be subject to human review. Member States shall also ensure that out-of-court redress mechanisms are available for the settlement of disputes. Such mechanisms shall enable disputes to be settled impartially and shall not deprive the user of the legal protection afforded by national law, without prejudice to the rights of users to have recourse to efficient judicial remedies. In particular, Member States shall ensure that users have access to a court or another relevant judicial authority to assert the use of an exception or limitation to copyright and related rights.

This Directive shall in no way affect legitimate uses, such as uses under exceptions or limitations provided for in Union law, and shall not lead to any identification of individual users nor to the processing of personal data, except in accordance with Directive 2002/58/EC and Regulation (EU) 2016/679.

Online content-sharing service providers shall inform their users in their terms and conditions that they can use works and other subject matter under exceptions or limitations to copyright and related rights provided for in Union law".

[36] João Pedro Quintais et al., *Safeguarding User Freedoms in Implementing Article 17 of the Copyright in the Digital Single Market Directive* (JIPITEC 10(3), 2019) 277–282.

[37] Krzysztof Garstka, Guiding the Blind Bloodhonds: How to mitigate the risks Article 17 of Directive 2019/970 poses to the freedom of expression in: in: Paul Torremans (eds.) *Intellectual Property and Human Rights* (Wolters Kluwer Law & Business 2020) 335.

risk of broad limitations to lawful content due to AI's lack of contextual sensitivity to detect exceptions and limitations.

3.1 Automated preventive filtering and risk of over-blocking

There is an internal conflict within the systematic structure of Article 17. Specifically, par. (7) provides that the cooperation between rights holders and Intermediaries – presented in par. (4) – shall not prevent *ex ante* the availability of content uploaded by users which does not infringe copyright including, especially if it is covered by an exception or limitation.[38] At the same time, par. (4)(b) encourages Intermediaries to make preventive "*best efforts*" to ensure the unavailability of specific works, in order to avoid secondary liability. Here originates the issue from the use of AI-based algorithms for preventive filtering: the obligation to ensure that users can upload lawful content, while preventing copyright infringing uploads, is extremely difficult to realise with automated means, especially in cases of context-contingent uses under copyright exceptions or limitations. Things are more complicated when trying to program into an AI system the hierarchy between Article 17(7), formulated as an absolute standard (*"shall not result in the prevention of the availability of works or other subject matter uploaded by users"*), and 17(4), which is based on a relative criterion such as the "*best efforts*" obligation to obtain authorisation or make the content unavailable expeditiously.

As a consequence of regulatory and stakeholders' pressure on Intermediaries to provide an easier mechanism towards infringing content removal, and since the coming into force of the DMS Directive in 2019, which affected the monitoring obligations of Intermediaries, the Intermediaries have largely implemented AI-based automated preventive filtering and blocking of content at the point of upload, before it is even made available to the public.[39] This general, algorithmic filtering (that leverages machine learning to restrict upload *ex ante*) is difficult to justify as it might result in over-blocking of lawful uses of content. As seen fur-

[38] See very clearly in this sense, with references to the protection of the fundamental rights of users, Recital 70 DSM Directive.

[39] Martin Senftleben, *Institutionalized Algorithmic Enforcement – The Pros and Cons of the EU Approach to UGC Platform Liability* (Florida International University Law Review 14 2020) 299–328.

ther, it might even account as a form of censorship since it could cause disproportionate consequences and detrimental effects on users' freedoms in comparison with the protection of copyright holders required by Article 17.

Similarly, in its recent Guidance on Article 17, the EU Commission has shifted from a position that rejected *ex ante* blocking of content to a more permissive take towards *ex ante* blocking beyond manifestly illegal content.[40] By allowing rights holders to 'earmark' content "*unauthorised online availability of which could cause significant economic harm to them*"[41] they can circumvent the principle that automatic blocking should be limited only to manifestly infringing uses. Consequently, uploads that include 'earmarked' protected content do not benefit from the *ex ante* protections for likely legitimate uses, allowing Intermediaries to use AI-based filters to block its upload from the beginning. Thus, the Guidance promotes a switch to a system based on privately governed mechanisms of preventive monitoring and enforcement of automated filtering, which undermines the principle that automated filtering cannot limit lawful upload and overcome the use of exceptions and limitations.

This approach of the EU legislator that sees automated algorithmic filtering as a necessary consequence for Intermediaries to discharge their monitoring obligations, even if in combination with other non-automated mechanisms,[42] is based on the misconception that AI might be able to solve all copyright enforcement problems. Instead, in the current state of the technology, algorithmic content moderation is not as sophisticated as believed. It is best to keep in mind how extremely difficult it is to programme into AI-based automated systems all contextual factors needed to be assessed to avoid overenforcement by filtering, as shown by

[40] Communication from the Commission to the European Parliament and the Council: Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market, COM/2021/288 Final ('Guidance').

[41] "*When providing the relevant and necessary information to the service providers, rightholders may choose to identify specific content which is protected by copyright and related rights, the unauthorised online availability of which could cause significant economic harm to them. The prior earmarking by rightholders of such content may be a factor to be taken into account when assessing whether online content-sharing service providers have made their best efforts to ensure the unavailability of this specific content and whether they have done so in compliance with the safeguards for legitimate uses under Article 17(7), as explained in part VI below*" Guidance, section V.2. p.14.

[42] Gerald Spindler, *The Liability system of Art. 17 DSMD and national implementation – contravening prohibition of general monitoring duties?* (JIPITEC 10 2020) 356.

recent empirical studies of automated copyright enforcement that report substantial over-blocking of content on video sharing platforms.[43] In any case, although Article 17(8) is very clear in stating that the fulfilment of the Intermediaries' obligations shall not lead to a general monitoring obligation. Intermediaries should not presume the infringing nature of contents. Thus, the availability of uploaded content for the public should be limited by fully automated filtering only for cases of manifestly infringing uploads (i.e. material that is identical or equivalent to the 'earmarked' content, previously requested by the rights holders).

3.2 Automated preventive filtering and lack of contextual sensitivity to detect exceptions and limitations

When fulfilling their monitoring obligations, Intermediaries must be careful not to allow algorithms to restrict users' rights to lawfully share and access information. Recital 70 DSM Directive explicitly recognises the importance of striking a balance between the right to intellectual property (Article 17(2)) and the fundamental freedom of expression and freedom of the arts, respectively, under Articles 11 and 13 of the Charter of fundamental rights of the EU[44]. By introducing this balance, the EU legislator has decided to award special status to these new mandatory exceptions and limitations, grounding their basis in fundamental rights.[45]

Achieving this balance between different fundamental freedoms and rights is largely dependent on the technological solutions that Intermediaries will employ to discharge their obligations. Without further repetition, it is worth underlining here the potential conflicts between the required monitoring obligations and the risk of misidentification and over-blocking when using algorithms for content moderation – due to their lack of contextual sensitivity in detecting specific context-related

[43] See e.g. Sharon Bar-Ziv, Niva Elkin-Koren, *Behind the scenes of online copyright enforcement: Empirical evidence on notice & takedown* (Connecticut Law Review, Vol. 50, 2017); or Kris Erickson and Martin Kretschmer, This video is unavailable (JIPITEC 9(1)2018).

[44] Article 11 of the Charter on freedom of expression and information states: "1. *Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. 2. The freedom and pluralism of the media shall be respected.*"

Article 13 of the Charter on Freedom of the arts and sciences affirms: "*The arts and scientific research shall be free of constraint. Academic freedom shall be respected.*"

[45] See e.g. op. cit. Emma Llansò et al. (2019).

elements. In fact, automated, AI-based filters are unable to recognise contextual nuances, which are necessary to distinguish *prima facie* infringements from uses that fall within the scope of exceptions or limitations provided by the law (e.g. reproduction of a part of a work for parodic use or permitted quotation).

A similar approach is portrayed in the EU Advocate General's opinion in the recent case C-401/19, through which the Polish Government has filed an action for annulment of Article 17 due to violation of freedom of expression under Article 11 of the EU Charter. According to previous CJEU case law which rejected general monitoring obligations that would monitor all the transmissions within a network[46], the AG describes how the 'generality' of an obligation will not have to be determined by the amount of information processed, but by the specific content that is being surveyed. He then illustrates how this provision actually imposes a 'specific' monitoring obligation to 'ensure the unavailability of specific works and other subject matter' previously earmarked by the rights holders ex Article 17(4). Any other conclusion (such as considering this a 'general' monitoring obligation) "*de facto obliges an intermediary provider to filter, using software tools, all of the information uploaded by the users of its service, even if it is a matter of searching for specific infringements, (and it) would regrettably amount to ignoring the technological developments which make such filtering possible and to depriving the EU legislature of a useful means of combating certain types of illegal content*"[47].

In light of the above, whether these concerns can be mitigated with effective and appropriate technological measures will be decisive in combatting unduly restrictions of fundamental freedoms.[48] Quintais et al. note that the application of preventive algorithmic content moderation is only possible as long as a proper filtering technology is available on the market and meets the legal requirements set forth in Article 17. In essence, preventive algorithmic filtering should only be allowed if it: (i)

[46] *Scarlet Extended SA v. Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)*, C-70/10, ECLI:EU:C:2011:771; *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV*, C-360/10, ECLI:EU:C:2012:85; *Tobias Mc Fadden v, Sony Music Entertainment Germany GmbH*, C-484/14, ECLI:EU:C:2016:689.

[47] Advocate General's Opinion in Case C-401/19, Poland v Parliament and Council, 15 July 2021, point 113.

[48] Sere e.g. Christophe Geiger and Bernd Justin Jütte, *Platform liability under Article 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match* (GRUR International, Vol 70, 2021).

meets the proportionality requirements in paragraph 17(5); (ii) enables the recognition of the mandatory exceptions and limitations in paragraph 17(7), including their contextual and dynamic aspects; and (iii) in no way affects legitimate uses, as mandated in paragraph 17(7) and (9).[49]

3.3 The future legal framework on Intermediaries' liability

As seen above, Article 17 is an extremely complex legal provision. As Dusollier notes, it is the "*monster provision of the Directive, both by its size and hazardousness*"[50]. The difficulties in interpreting the provision, and for Intermediaries to apply it correctly, in order to exclude liability is shown by the significant legal scholarship already existing, most of which was written even before the national implementation deadline[51].

To complicate things, on 15 December 2020, the EU Commission proposed two legislative initiatives to upgrade rules governing digital services in the EU to create a safer and more open digital space "*in which the fundamental rights of all users of digital services are protected*"[52]. The first of these proposals introduces Regulation on a Single Market for Digital Services (Digital Service Act), which provides a new regulatory approach to online Intermediaries through horizontal rules interlacing with a variety of EU legislations. For the purpose of this paper, it is relevant to highlight how the overlap between the Digital Service Act and the DSM Directive will shape the future legal framework around Intermediaries' liability and will impact how the latter shall programme or update their algorithms for automated filtering and content moderation, in order to enjoy the relevant liability exemptions.

[49] Op. cit. João Pedro Quintais, et al. (2019).

[50] Séverine Dusollier, *The 2019 Directive on copyright in the digital single market: some progress, a few bad choices, and an overall failed ambition* (Common Market Law Review, Vol. 57 2020).

[51] For a compilation of interventions and publications see: <create.ac.uk/cdsm-implementation-resource-page/#consultations-transpositions> accessed 7 August 2021. At the time of writing, the implementation of the DSM Directive at national level has been quite slow with only two Member States having fully completed the transposition into national law, partially also due to the discussions and uncertainty involving this provision.

[52] Proposal for a regulation of the European Parliament and of the Council on a Single Market For Digital Services ('Digital Services Act') and amending Directive 2000/31/EC, COM/2020/825 final.

First of all, both the DSM Directive and the proposed Digital Services Act establish obligations on various online intermediaries (including those that are central to the argument of this paper, i.e. online content sharing service providers) on how to handle illegal information. The DSM Directive targets copyright infringing content and the Digital Services Act targets illegal content in general (including content which infringes on copyright). Despite the fact that the two instruments have a different legal nature (the DSM Directive will have to be transposed in Member State law, whereas the Digital Services Act is a directly applicable Regulation), and that they seem to operate at different, complementary levels, it is worth mentioning potential overlaps between the two legislations and present how the Intermediaries' liability framework might look like in the future, if and when the Digital Service Act might enter into force.

At first sight, these regimes do not particularly overlap with each other since Article 17 is *lex specialis*, as per Recital 9 and Article 1(5)(c) Digital Service Act. The latter states that the Regulation is "*without prejudice to the rules laid down by (…) Union law on copyright and related rights*"[53]. However, in the view of certain scholars, this unaffected result "*can only relate to aspects which indeed are specifically covered by those* (copyright) *rules*"[54]. In fact, the intersection between the DSM Directive and the Digital Service Act is more complex than what Recital 9 Digital Service Act and Article 1(5)(c) Digital Service Act seems to suggest at first sight. The EU Commission provided to the Council's Working Party on Intellectual Property (Copyright) some internal insights on how the relationship between the two instruments can be interpreted.[55] The Commission

[53] Recital 9 Digital Service Act states that "*This Regulation should complement, yet not affect the application of rules resulting from other acts of Union law regulating certain aspects of the provision of intermediary services […]. Therefore, this Regulation leaves those other acts, which are to be considered lex specialist in relation to the generally applicable framework set out in this Regulation, unaffected. However, the rules of this Regulation apply in respect of issues that are not or not fully addressed by those other acts as well as issues on which those other acts leave Member States the possibility of adopting certain measures at national level"*. Supporting Recital 11 Digital Service Act adds that the "*Regulation is without prejudice to the rules of Union law on copyright and related rights, which establish specific rules and procedures that should remain unaffected*".

[54] João Pedro Quintais et al., Interim report on mapping of EU legal framework and intermediaries' practices on copyright content moderation and removal, ReCreating Europe (2021), p. 43.

[55] Council of the European Union, Working Paper, N° Cion doc.: 14124/20, Digital Services Act and EU copyright legislation – Information from the Commission (2021).

has advised that the "*Digital Service Act is not an IPR enforcement tool*" given its general and horizontal nature. Nevertheless, it "*includes a full toolbox which can be very useful for the enforcement of IPR*" and should be applied "*without prejudice to existing IPR rules*". In short, it looks like the Commission upholds that Article 17 DSM Directive will remain "*unaffected*" by the rules on liability proposed in the Digital Service Act. Only time will tell if that will be the case. As of now, without digressing too much on this topic, the Digital Service Act is believed to apply to Intermediaries only insofar as it contains rules that regulate matters not covered by Article 17 DSM Directive, or in cases of specific matters which Article 17 leaves to the discretion of Member States.[56]

4 Conclusions

As seen above, due to the pressures from governments and users alike to take a more active role, more and more Intermediaries are turning towards AI to moderate online content on a large scale, since the deployment of AI-based algorithms can give rise to successful, automated detection, evaluation and removal of infringing content.

Accordingly, the EU legislator has increasingly become more accepting of the idea of imposing preventive filtering obligations on Intermediaries to screen out copyright infringing content and to hold Intermediaries accountable for copyright infringements originating from their users' activity. This, thanks to the assumption that algorithmic filtering technologies have become more sophisticated.

Having presented the legal framework and the interpretative issues around Article 17, I showed how AI algorithms play a fundamental role in helping fight online copyright infringements thanks to a 'predict and prevent' approach: considering the immense amount of data generated daily by users, these algorithms can quickly analyse huge amounts of data, proactively evaluate if a work contains infringing content, and then remove it.

Nevertheless, this approach, which sees automated algorithmic filtering as a necessary consequence for Intermediaries to discharge their mon-

[56] This includes e.g. rules from the Digital Service Act relating to the liability and to due diligence obligations for online Intermediaries of different sizes. For a speculative interpretation see: João Pedro Quintais, Sebastian, Felix Schwemer, *The interplay between the digital services act and sector regulation: how special is copyright*? (Forthcoming 2021).

itoring obligations, is based on the misconception that AI technology might be able to solve all copyright enforcement problems. Instead, in the current state, AI technology is not as sophisticated as believed and surely not a panacea for all issues related to online infringements. AI is still characterised by many technical limitations which conflict with the protection of users' fundamental rights and freedom. First, it might present serious danger of over-blocking lawful contents, unduly restricting freedoms as collateral effect of the 'block-first, verify-later' type of approach. Secondly, AI lacks contextual sensitivity, necessary to distinguish *prima facie* infringements from uses that fall within the scope of exceptions or limitations provided by the law. Whether these concerns can be mitigated with effective and appropriate technological measures will be decisive in combatting unduly restrictions of fundamental rights and freedoms by AI-based filtering algorithms.

In light of the above, the legislators (both at EU and national level) should therefore act with caution and avoid narratives about all-powerful algorithms, instead helping to shape users' online experience through provisions that combine the deployment of AI filtering algorithms together with safeguards of fundamental rights and freedoms of users. It is thus paramount in the future to introduce standards to enhance transparency and accountability of content moderation practices (e.g. introducing secondary human-review, due diligence processes and other risk assessment methodologies), and to ensure that users have access to complaint and redress mechanisms, to remedy wrongly executed automated decisions by AI.

Mikael Hansson

Arbetsgivaren, artificiell intelligens och ansvaret

1 Inledning – avstamp i nya frågor

Den juridiska fakulteten har funnits sedan Uppsala universitet grundades 1477, arbetsrätten i nutida mening sedan mitten av 1800-talet ungefär. Det är alltså gamla företeelser, som nu ställs inför en så modern företeelse att den knappt finns ännu – artificiell intelligens, eller kort och gott AI. Tanken på att arbetet grundar sig på ett avtal i modern mening härrör från ett ideologiskt skifte i mitten på 1800-talet, då liberalismen ersatte en merkantil rättsordning. Ur det föds den moderna arbetsrätten, som alltså är uppbyggd kring en kontraktsrelation mellan två parter; de som kommit att kallas *arbetstagare* respektive *arbetsgivare*.

Det rättsideologiska skiftet i mitten på 1800-talet som lade grunden för den samtida arbetsrättsliga regleringen har samband med ett tekniskt skifte, ”den industriella revolutionen”. Det är svårt, åtminstone för en jurist, att analysera sin egen samtid i ett skede – epokskiften pekas enklast ut i efterhand. Vi kan dock tänka oss att vi befinner oss i en fjärde industriell revolution, efter ångkraften och mekaniseringen som gav industrisamhället, elektriciteten samt Internet.[1] Kanske kommer framtiden att tala om ”AI-revolutionen”, och tala om den tid som vi kallar nu. Den tanken kan framkalla rädsla för att ersättas, och att vissa arbeten försvin-

[1] Se Bruno Debaensts bidrag i föreliggande volym. Se också till exempel Lundqvist, Ulf, Artificiell Intelligens — rättsordning och rättstillämpning, SvJT 2020 s. 382–405 (s. 384 m.fl.).

ner på grund av teknisk utveckling är i och för sig givet.[2] En poäng, eller själva poängen, med att ersätta människa med maskin kan vara att maskinen är effektivare och billigare, ett annat (och relaterat) att den inte behöver omfattas av social skyddslagstiftning. Det är dock inte någon långsökt tanke att den ”AI-revolution” vi nu beskådar också skall driva fram en rättslig förändring. Tanken på att ett datorprogram, som utför arbete, skall vara en arbetstagare är dock om inte omöjlig så i vart fall onödig. AI behöver inte någon skyddslagstiftning, vare sig man betraktar arbetsrättslig skyddslagstiftning som ett utflöde av rättigheter och rättvisa eller som ett sätt att hindra att arbetskraften används på ett ohållbart sätt av kortsiktiga arbetsgivare. Tvärtom är åtminstone en del av poängen med att teknifiera arbetet att maskiner inte behöver vila, rekreation eller har något intresse av att umgås med kollegor i fikarummet. Frågor om social hållbarhet förlorar aktualitet om arbetet inte utförs av människor, i vart fall såvitt angår själva arbetet (att människor är arbetslösa är inte socialt hållbart, i vart fall inte om inte maskiner kan ta över till den grad att de försörjer alla). I det följande skall blicken istället vändas mot arbetsgivaren. En rädsla för att arbetstillfällen skall tas över av tekniken motsvaras inte av en motsvarande rädsla att tekniken skall ta över arbetsgivarens roll. I en framtidsspaning kan det ändå vara värt att ställa frågan, och frågan om arbetsgivarens identitet kan säga något om arbetsrättens identitet. Arbetsgivaren är, liksom arbetstagaren, ett rättsligt koncept. Undersökningen handlar om möjligheterna och begränsningarna i det rättsliga konceptet, samtidigt som den försöker undvika (om det är möjligt) att jaga bekräftelse på vår samtida förförståelse.[3] Studien görs i två delar – först diskuteras möjligheterna i arbetsgivarbegreppet förstått som

[2] Framtidsblickande samtidsskildringar med avstamp i teknisk utveckling finns det gott om, särskilt utanför den juridiska litteraturen; Bech, Ulrich, Risksamhället. På väg mot en annan modernitet, Daidalos 1986 behandlar industrisamhällets övergång till ett risksamhälle där rädsla för till exempel miljökatastrofer till följd av industrisamhället styr utvecklingen mot en (åtminstone delvis) ny era; Castells, Manuel, Informationsåldern. Ekonomi, samhälle och kultur beskriver industrisamhällets övergång till en ny epok i tre band (Nätverkssamhällets framväxt, 2 uppl. 2020, Identitetens makt 1998, Millenniets slut 2020), Daidalos; Susskind, Richard och Susskind, Daniel, Professionernas framtid. Hur teknologin kommer att förändra experters arbete, Daidalos 2017 diskuterar kunskapsbranschernas och dess framtid; Paulsen, Roland, Arbetssamhället. Hur arbetet överlevde teknologin, Atlas 2010 skriver om att människor trots teknologins framsteg och övertagande av vissa arbetsuppgifter arbetar mer än någonsin. Till exempel.

[3] Stenlund, Karolina, Rättighetsargumentet i skadeståndsrätten, Iustus 2021 s. 56.

en invertering av arbetstagarbegreppet, därefter görs nedslag i tre rättsfall som handlar om ansvar för diskriminering.

2 Arbetsgivarens identitet – ”arbetsgivarbegreppet”

Arbetsrättsliga reglers tillämplighet avgörs av det så kallade *arbetstagarbegreppet*. Att tala om ett begrepp som rättsfaktum på det sätt som det etablerade språkbruket gör är nog lite oegentligt, men alla vet att fråga är om tillämpligheten av ett visst rekvisit i en viss regel; 1 § semesterlagen (1977:480), 1 kap. 2 § arbetsmiljölagen (1977:1160), 1 § lagen (1982:80) om anställningsskydd, 1 § arbetstidslagen (1982:673), 1 § lag (1994:260) om offentlig anställning, 1 § lag (2012:854) om uthyrning av arbetstagare, och så vidare. Vissa lagar innehåller också en utvidgning av tillämpningen, till exempel 6 kap. 5 § skadeståndslagen (1972:207) och 1 § 2 st. lag (1976:580) om medbestämmande i arbetslivet (medbestämmandelagen), men utvidgningen utgår då från, och bekräftar därmed, det civilrättsliga arbetstagarbegreppet.[4] Det finns en omfattande praxis och en omfattande litteratur som tar sikte på *arbetstagarens* identitet.[5] Frågan avgörs genom ett antal kriterier, varav ett antal betraktas som grundläggande rekvisit; arbetstagaren är en *fysisk person*, som utför *arbete*, *för annans räkning*, på grund av *avtal*. Därutöver brukar ett antal bedömningskriterier tillämpas, som kan vara mer eller mindre relevanta i det enskilda fallet beroende på arbetets karaktär. Medan ”rekvisiten” måste föreligga ligger bedömningskriterierna till grund för en helhetsbedömning. Bedömningen siktar på att avgöra om en viss person (som alltså måste vara en fysisk person) är arbetstagare (det vill säger uppfyller det

[4] Svante Bergström skiljer på allmänna skillnader mellan olika arbetstagarbegrepp såsom ett socialt och ett civilrättsligt arbetstagarbegrepp (skillnader som inte framgår uttryckligt i lagarna), och speciella skillnader mellan enstaka lagars arbetstagarbegrepp (skillnader som framgår av speciella bestämmelser i respektive lag), Bergström, Svante, Kollektivavtalslagen. Studier över dess huvudprinciper, Uppsala 1948 s. 33.

[5] Se Selberg, Niklas, Arbetsgivarbegreppet och arbetsrättsligt ansvar i komplexa arbetsorganisationer. En studie av anställningsskydd, diskriminering och arbetsmiljö. Juridiska fakulteten, Lunds universitet 2017 s. 27 ff. Som den första utvecklade studien anförs Adlercreutz, Axel, Arbetstagarbegreppet, P.A. Norstedt & söners förlag 1964. Adlercreutz (s. 13) hänvisar i sin tur till den schematiska uppställningen i Bergström, Kollektivavtalslagen s. 39 f.

rekvisitet i den åberopade regeln) eller något annat (den regel som då inte kan tillämpas bryr sig inte om vad). Arbetstagaren utför arbete under arbetsgivarens ledning och kontroll, medan uppdragstagaren själv planerar och kontrollerar arbetet. Arbetstagaren står inte själv för material, verktyg och liknande, medan uppdragstagaren använder egna verktyg. Arbetstagaren får ersättning för utlägg, medan uppdragstagaren räknar in det i priset för tjänsten. Arbetstagaren får ersättning (”lön”) baserad på arbetad tid snarare än efter prestation, medan uppdragstagaren får ersättning efter utfört arbete. Arbetstagarförhållandet är som utgångspunkt varaktigt (anställningsavtal anses *in dubio* ingångna på obestämd tid, se 4 § lagen om anställningsskydd), medan det idealtypiska uppdragsförhållandet gäller en viss prestation eller ett visst uppdrag. Arbetstagaren kan inte sätta någon annan i sitt ställe, medan uppdragstagarens prestation är knuten till resultatet snarare än vem som utför den. Arbetstagaren är en del av arbetsgivarens organisation, medan uppdragstagaren är självständig i förhållande till denna. Och så vidare.

Bedömningen är alltså inriktad på arbetstagaren, medan arbetsgivaren kort och gott brukar definieras som arbetstagarens motpart i kontraktet.[6] Definitionen av om någon är arbetsgivare eller uppdragsgivare (eller vilken beteckningen nu är) följer alltså av definitionen av motparten. Frågans orientering har att göra med att det arbetsrättsliga regelverket är en skyddslagstiftning som ger arbetstagaren rättigheter gentemot arbetsgivaren, och det är svårt att se fallet då den arbetspresterande parten har intresse av att i motsats till motparten hävda att hen inte är arbetstagare utan uppdragstagare. Däremot kan tänkas att arbetstagaren har intresse av att en viss person är arbetsgivare och inte en annan, det vill säga vem av två eller flera möjliga subjekt som är arbetsgivare. Det är enligt Källström och Malmberg den vanligaste frågan när man talar om arbetsgivarbegreppet. De andra frågorna är för vem arbetsgivaren ansvarar (principalansvar) och begränsningar i arbetsgivarens skyldigheter i olika avseenden (till exempel omplaceringsskyldighetens omfattning enligt 7 § 2 st. lagen om anställningsskydd, som begränsar omplaceringsskyldigheten till att arbetsgivaren

[6] Jfr Selberg, Arbetsgivarbegreppet och arbetsrättsligt ansvar i komplexa arbetsorganisationer s. 47 ff. Se också till exempel Källström, Kent och Malmberg, Jonas, Anställningsförhållandet, 4 uppl. Iustus 2016 s. 40; Glavå, Mats och Hansson, Mikael, Arbetsrätt, 4 uppl. Studentlitteratur 2020 s. 87.

är skyldig att bereda annat arbete *hos sig*).[7] Ingen av de tre situationerna rymmer dock frågan om det finns en arbetsgivare i och för sig.

Eftersom arbetsgivaren är motparten till arbetstagaren skulle arbetsgivarbegreppet kunna formuleras genom en spegling av de kriterier som definierar arbetstagaren. Arbetsgivaren behöver inte vara en *fysisk person*, tvärtom är det vanligt att arbetsgivaren är en juridisk person. Vilka krav som ställs på en juridisk person är ett problem som arbetsrätten inte brukar göra till sitt. Det grundläggande kravet på att arbetstagaren skall vara en fysisk person brukar ta sikte på att juridiska personer (det vill säga föreningar, bolag, stiftelser och så vidare) inte kan uppträda som arbetstagarpart i ett anställningsavtal. Om det skall förstås som ett positivt krav på arbetsgivaren (arbetsgivaren måste vara en fysisk *eller* juridisk person) kan inte den som inte kan uppnå rättskapacitet som fysisk eller juridisk person vara part i ett anställningsavtal. Ställs frågan om AI eller någon annan process kan bära förpliktelser i ett anställningsavtal, förflyttar sig bedömningen till den associationsrättsliga frågan om processen kan bilda eller ingå i bolag. Om kravet istället skall förstås negativt (arbetsgivaren behöver inte ha en viss rättspersonlighet) förflyttar spörsmålet sig till den avtalsrättsliga frågan om AI kan ingå avtal. I båda fallen har frågan med *rättskapacitet* att göra. Svaret är antagligen nekande, men den indelning i rättsområden som arbetstagarbegreppet (och därmed arbetsgivarbegreppet) är en del av får till följd att synfältet begränsas. Arbetsrätten har inte svaret – och utesluter alltså inte ansvar enligt arbetsrättsliga regler för AI.

Det leder in på ett annat grundläggande krav, nämligen att arbetet skall utföras på grund av *avtal*. Om det kravet förstås positivt utgår det från att såväl arbetsgivare och arbetstagare måste vara avtalsparter, alltså att de kan tillskrivas rättskapacitet. Om kravet istället förstås negativt handlar det dock om att skära bort visst typ av arbete från arbetsrätten och de arbetsrättsliga reglerna, till exempel arbete som är en del av utbildning, arbete till följd av värnplikt eller inom ramen för kriminalvård. Grunden är då inte ”frivillighet” på det sätt som förutsätts till följd av avtal, utan till följd av ett visst (och annan form av, ett offentligrättsligt) tvång. Med den förståelsen är dock inte avtalet i sig avgörande, utan poängen är att avskära vissa företeelser från vissa regler. Det utesluter dock inte att en relation faller under arbetstagarbegreppet, även om den ena

[7] Källström och Malmberg, Anställningsförhållandet s. 40 ff. Se också Adlercreutz, Arbetsgivarbegreppet s. 162 ff.

parten inte har rättskapacitet som avtalspart. Inte heller det utesluter i så fall ansvar enligt arbetsrättsliga regler för AI.

För att den arbetspresterande parten skall vara arbetstagare krävs att arbetet utförs för *annans räkning*.[8] En fråga kan då vara om den som mottar prestationen därmed är arbetsgivare, eller om kravet tar sikte på att någon annan än den arbetspresterande tillgodogör sig arbetets värde. I bemanningsarbete ligger det i själva konstruktionen att arbetet utförs hos någon annan än den som är arbetsgivare i kontraktuell mening.[9] I lagen om uthyrning av arbetstagare används dock inte begreppet arbetsgivare, utan den som har arbetsgivarskyldigheterna är ett ”bemanningsföretag” (1 och 5 §§). Skall kravet på att arbete skall utföras för annans räkning förstås positivt så att den som kan tillgodogöra sig värdet av arbete är arbetsgivare anknyter det till den ovan berörda förutsättningen att motparten till arbetstagaren måste ha rättskapacitet. Om kravet däremot förstås negativt, så att det utesluter att reglerna tillämpas när den som utför arbetet själv är den som drar nytta av det är det snarare en förlängning av kravet på avtal. Negativt förstått blir dock innebörden inte att det måste finnas ett rättsligt giltigt avtal enligt avtalsrättsliga regler, utan att arbetet skall utföras i en relation med någon annan. Poängen med ett negativt kriterium för arbetsgivarbegreppet i det avseendet blir att den som utför arbete för egen räkning inte behöver följa arbetsmiljö- eller arbetstidsregler och inte ge sig själv semester. Det spelar dock i så fall ingen avgörande roll vem någon annan är eller dennes rättsliga status, och ansvar enligt arbetsrättsliga regler utesluts inte.

Vidare brukar, som framgått, den rättsliga bedömningen av arbetstagarbegreppet göras utifrån ett antal bedömningskriterier.[10] Ett centralt sådant är att arbetsgivaren utövar ledning och kontroll över arbetet. Sådana ledningsfunktioner kan säkert utföras av AI; fördela körningar

[8] Adlercreutz behandlar i Arbetstagarbegreppet frågan om arbetet utförs av annan under rubriken *Vem är att anse som arbetsgivare? Mellanmanssituationen m.m.* Jämfört med de AI-tillämpningar som är föremålet för den föreliggande volymen är tiderna i Adlercreutz undersökning andra – skogskörare återkommer (se till exempel s. 178, s. 185, s. 464 och s. 467) och betskötsel förekommer (s. 210).

[9] Bemanningsarbete blev tillåtet på den svenska arbetsmarknaden 1992, varför äldre studier om arbetstagarbegreppet inte gör sig det problemet. Se Berg, Annika, Bemanningsarbete, flexibilitet och likabehandling. En studie av svensk rätt och kollektivavtalsreglering med komparativa inslag, Juristförlaget i Lund s. 15 ff. och s. 106 ff.

[10] Se ovan och till exempel sammanställningar i Källström och Malmberg, Anställningsförhållandet s. 26 f. och Glavå och Hansson, Arbetsrätt s. 84.

smart, övervaka produktion och så vidare. Inriktas bedömningen på det faktiska förhållandet, med bortseende från arbetsgivarpartens rättskapacitet finns inget som hindrar att det kriteriet pekar på AI eller något annat som arbetsgivare. Andra bedömningskriterier har ingen tydlig bäring på parternas identitet eller rättsliga status; om maskiner och arbetsmaterial tillhandahålls av den som utför arbetet (vilket talar mot ett arbetstagarförhållande) eller av den som tillgodogör sig det, om den som utför arbete får ersättning för utlägg (vilket talar för ett arbetstagarförhållande eller inte) eller inte. Åter andra bedömningskriterier, såsom om relationen är varaktig eller om den arbetspresterande är beroende av en motpart knyts mer eller mindre uttryckligt till avtalsförhållandet. Det finns dock i och för sig ingen logiskt nödvändig koppling till ett avtalsförhållande där arbetsgivarparten har en viss rättslig status. Bedömningskriterierna är utmejslade genom ett drygt sekels rättspraxis och sammanställda i rättsvetenskapen – allt med ett avtalsförhållande mellan en fysisk person och en juridisk (eller fysisk) person för ögonen. Förekomsten av AI eller digitala processer har inte varit i blickfånget, men att utifrån det dra slutsatsen att reglerna skulle utesluta sådana företeelser låter sig inte göras – och det finns inte heller någon anledning att utforma regler för att utesluta något som inte finns.

3 ”Arbetsgivarens” ansvar – tre nedslag

I det följande skall tre fall beröras, som handlar om arbetsgivarens identitet. Eller rättare, i två av tre fall en *presumtiv* arbetsgivares ansvar eftersom de handlar om personer som av olika skäl sorterats bort från anställningsprocesser. Om fallen skall sättas in i den systematik för arbetsgivarbegreppet som presenteras av Källström och Malmberg sorteras de in under den andra kategorin, det vill säga frågan om för vilket handlande en arbetsgivare ansvarar (se ovan vid not 7). Ovan var frågan om arbetsgivare kan göras ansvariga som sådana, i de nedan diskuterade fallen är frågan om arbetsgivaren kan undgå ansvar. Inget av fallen rör AI eller automatiserade processer i och för sig, utan frågan är vilket ansvar eller vilka möjligheter att freda sig från ansvar en arbetsgivare har.

AD 2007 nr 45 – Laika film

Det finns, som sagt, ganska lite praxis som behandlar arbetsgivarbegreppet. Från den tiden då den äldre diskrimineringslagstiftningen, som ersattes av diskrimineringslagen (2009:567), märks AD 2007 nr 45. Diskrimineringslagen syftade till att samla de diskrimineringsförbud som var spridda på olika lagar. Genom diskrimineringslagen slogs sju civilrättsliga lagar ihop till en, samtidigt som fyra ombudsmän förenades i en. Diskrimineringslagen syftade också till att genomföra ett antal direktiv på diskrimineringsområdet (se nedan om det så kallade arbetslivsdirektivet). I delar där diskrimineringsförbud grundade i EU-lagstiftning fördes över från de äldre lagarna till diskrimineringslagen är äldre rättspraxis alltjämt relevant.

AD 2007 nr 45 rörde frågan om ansvar för direkt diskriminering enligt dåvarande lagen (1999:130) om åtgärder mot diskriminering i arbetslivet på grund av etnisk tillhörighet, religion eller annan trosuppfattning. En arbetstagare i underordnad ställning hade sorterat bort en ansökan på grund av att den innehöll många stavfel, och meddelat sökanden att de sökte någon som behärskade svenska. Såvitt var visat i målet hade arbetstagaren handlat på eget initiativ, bolagets behöriga företrädare hade inte instruerat arbetstagaren att sortera bort någon ansökan och de kände inte heller till tilltaget. Att förfarandet i och för sig utgjorde direkt diskriminering stod klart (svarandebolaget vitsordade att meddelandet var diskriminerande), däremot stod det inte klart om arbetsgivaren skulle ansvara för diskrimineringen. Frågan hängs upp på rekvisitet *arbetsgivare* i 10 § 1999 års lag.[11] Den lagens konstruktion av diskrimineringsförbuden var väsentligen densamma som i diskrimineringslagen; i 8–9 b §§ definierades diskriminering i form av direkt diskriminering, indirekt diskriminering, trakasserier samt instruktioner att diskriminera, och i 10 § angavs när det var förbjudet att diskriminera. Såvitt relevant i målet löd 10 §: Förbuden i 8–9 b §§ gäller när arbetsgivaren 1. beslutar i en anställningsfråga, tar ut en arbetssökande till anställningsintervju eller vidtar annan åtgärd under anställningsförfarandet.

I domskälen läggs fokus på frågan om förbudet är tillämpligt. Det är samtidigt en fokusförskjutning från frågan om arbetsgivaren, det vill säga svarandebolaget, skall ansvara för den i och för sig diskriminerande handlingen, till frågan om ”arbetsgivarbegreppet”. Glidningen är alltså densamma som den inledningsvis berörda, att ”arbetsgivarbegreppet” är

[11] I fråga om innehållet i det rekvisitet motsvarar de då gällande reglerna den nu gällande diskrimineringslagen, se Hellborg, Sabina, Diskrimineringsansvar. En civilrättslig undersökning av förutsättningarna för ansvar och ersättning vid diskriminering, Iustus 2018 s. 140 not 512.

en tillämpning av rekvisitet ”arbetsgivare” i en viss bestämmelse. Den innebär dock att frågan inte ställs ”skall förbudet mot diskriminering tillämpas och skadestånd utdömas?” (vilket vore att svara på kärandens yrkande) utan så att den gäller innehållet i arbetsgivarbegreppet. Frågan blir om arbetsgivaren är arbetsgivare i relevant mening och att bolaget därför skall ansvara för den i och för sig konstaterade diskrimineringen (vilket snarare svarar på en inte förd fastställelsetalan). Svaret söks i förarbetena, och i 22 § jämställdhetslagen. Resonemangen i AD 2007 nr 45 hämtas från AD 2007 nr 16 som också rörde frågan om en arbetsgivares ansvar för andras handlande, men i det fallet för fackliga företrädares diskriminerande yttranden under en anställningsintervju (det kan noteras att Arbetsdomstolens ordförande och sekreterare var desamma i de båda målen, som avgjordes med tre månaders mellanrum). I 22 § jämställdhetslagen, som innehöll ett repressalieförbud motsvarande 2 kap. 18 § diskrimineringslagen, fanns vad som av Arbetsdomstolen uppfattades som en precisering av arbetsgivarbegreppet: Den som i arbetsgivarens ställe har rätt att besluta om en arbetstagares arbetsförhållanden skall vid tillämpning av första och andra styckena likställas med arbetsgivare (samma ”precisering” finns där, liksom i 2 kap. 1 § diskrimineringslagen som motsvarar 10 § i 1999 års lag). Efter att ha konstaterat att en sådan precisering saknades i den tillämpliga lagen och att det i andra regler som ålade ”arbetsgivaren” ett ansvar (lagen (2003:307) om förbud mot diskriminering (som också upphävdes med införandet av diskrimineringslagen)) fanns ett uttryckligt principalansvar landade Arbetsdomstolen i att det saknades utrymme för att ge begreppet arbetsgivaren någon annan innebörd än vad som följer av annan arbetsrättslig lagstiftning, och att det alltså saknades utrymme enligt lagen för en mera vidsträckt tillämpning av arbetsgivarens ansvar.

Jonas Malmberg har kritiserat domskälen (med instämmande av Susanne Fransson och Eberhard Stüber), på den grunden att det inte är förenligt med EU-rätten att generellt undanta handlingar av arbetsgivaren från ansvar för diskriminering (se vidare nedan).[12] Här skall dock en annan fråga belysas. Glidningen i Arbetsdomstolens domskäl är inte bara en glidning från rekvisit till begrepp utan också från en syn på reglernas

[12] Malmberg, Jonas, Diskriminering och principalansvar, i Ahlberg, Kerstin (red.) Vänbok till Ronnie Eklund, Iustus 2010 s. 397–417 s. 411; Fransson, Susanne och Stüber, Eberhard, Diskrimineringslagen. En kommentar, 3 uppl. Norstedts Juridik 2021 s. 157. Se också Hellborg, Diskrimineringsansvar s. 142 f.

ändamål till en annan. Med en civilrättslig logik, till vilka diskrimineringsförbuden i arbetslivet brukar hänföras,[13] vore reglernas ändamål att fördela risken för den skada diskrimineringen inneburit, inklusive kränkningen. Rättsmedlet för att sanktionera överträdelser var också skadestånd. I diskrimineringslagen har skadeståndet bytts mot diskrimineringsersättning, vilket innehåller ett större mått av straff.[14] Civilrätten fördelar risker mellan olika subjekt med olika intressen, medan *skuld* placeras på ett utpekat subjekt av straffrättsliga regler.[15] Om skuld skall åläggas följer det av legalitetsprincipen att tillämpningen skall vara snäv och att skuld inte skall åläggas utöver vad regeln tydligt anger. Det följer också att en person inte skall åläggas skuld för någon annans handlande med mindre att det klart framgår att det finns ett ansvar för dennes handlande. Begränsningen gäller så länge det mellankommande handlandet är en människas – om en automatiserad, icke-mänsklig, process skulle sortera ut ansökningar skulle det rimligen inte komma på fråga att fria den arbetsgivare som implementerat processen från ansvar för diskriminering på den grunden att diskrimineringen inte kunde tillräknas denne (frågan om köprättsligt ansvar för den som levererat systemet är en annan). Att processen skulle kunna sägas agera på eget bevåg gör därvid ingen skillnad – skulden, och därmed det rättsliga ansvaret, bärs av en människa. Där det saknas utrymme för att utsträcka ett begrepp så att ansvaret för en människas skuld bärs av någon annan, skulle det saknas utrymme för att undanta en människas ansvar för en automatiserad process.

[13] Se till exempel prop. 2007/08:95 Ett starkare skydd mot diskriminering s. 1 och s. 80.

[14] Se särskilt Hellborg, Diskrimineringsansvar s. 344 ff. Hellborg visar att frågan om skadestånd eller straff bara kan upprätthållas i vissa situationer och att nyanserna snarare finns i att diskrimineringslersättningen är både och (s. 377), men här får den grovyxade uppdelningen tjäna som retorisk figur (till mitt försvar underkänner inte Hellborg distinktionen som sådan).

[15] Se Andersson, Håkan, Ansvarsproblem i skadeståndsrätten. Skadeståndsrättsliga utvecklingslinjer. Bok I, Iustus 2013 s. 61 ff.; Asp, Petter, Ulväng, Magnus och Jareborg, Nils, Kriminalrättens grunder, 2 uppl. Iustus 2013 (omtryckt 2020) s. 58 f. och s. 269 ff.; Asp, Petter och Ulväng, Magnus, Straffrätt. En kortfattad översikt, 2 uppl. Iustus 2019 s. 47. Se också Fridström-Montoya, Therése, Homo juridicus. Den kapabla människan i rätten, Iustus 2017 s. 89 ff.

AD 2013 nr 5 – Middag för alla, utom…

I AD 2013 nr 5 var frågan om arbetsgivaren gjort sig skyldig till föreningsrättskränkning. Föreningsrätten är på sätt och vis den mest grundläggande rättigheten för framförallt arbetstagarsidans organisering, med första tydliga uttryck i den så kallade decemberkompromissen av 1906. Då slog de ganska nybildade organisationerna Svenska Arbetsgivarföreningen (SAF, numera Svenskt Näringsliv) och Landsorganisationen (LO) i avtal fast att föreningsrätten å ömse sidor skulle lämnas okränkt. Även om rättigheterna alltså riktar sig till båda sidor i det arbetsrättsliga avtalsförhållandet är det arbetarnas intresse av att sluta sig samman i fackföreningar som motiverar föreningsrätten. Den fackliga föreningsrätten har sedan lagfästs, nu i 7–9 §§ medbestämmandelagen. Med föreningsrätt avses enligt 7 § medbestämmandelagen rätt för arbetsgivare och arbetstagare att tillhöra arbetsgivar- eller arbetstagarorganisation, att utnyttja medlemskapet och att verka för organisationen eller för att sådan bildas. En kränkning av föreningsrätten föreligger enligt 8 § medbestämmandelagen om någon på arbetsgivar- eller arbetstagarsidan vidtager åtgärd till skada för någon på andra sidan för att denne har utnyttjat sin föreningsrätt ("repressaliefallet") eller om någon på ena sidan vidtager åtgärd mot någon på andra sidan i syfte att förmå denne att icke utnyttja sin föreningsrätt ("hindrandefallet"). Den fackliga föreningsrätten avskiljs från den allmänna friheten att sluta sig samman i föreningar genom sin placering i medbestämmandelagen, som gäller förhållandet mellan arbetsgivare och arbetstagare (se 1 § medbestämmandelagen).[16] I Europeiska konventionen om skydd för de mänskliga rättigheterna och de grundläggande friheterna (EKMR) är den fackliga föreningsrätten en del av den föreningsfriheten enligt art. 11, men det förtydligandet att den gäller också rätten att sluta sig samman i fackföreningar. Föreningsrätten enligt medbestämmandelagen gäller relationen mellan arbetsgivar- och arbetstagarsidan (inte motsättningar inom respektive sida), och reglerna slår till när en åtgärd utförs av någon. Ofta rör tvisterna i rättspraxis vad

[16] Prop. 1975/76:105 Bilaga 1 Lag om medbestämmande i arbetslivet s. 26.

som är en åtgärd eller vilket syfte den åtgärden har haft,[17] men här skall intresset istället riktas mot vem som är ”någon”.[18]

Ordalydelsen i 8 § medbestämmandelagen ger i det avseendet inte mer än att en kränkning av föreningsrätten föreligger när någon vidtager en åtgärd, och att denna någon skall befinna sig på ena sidan i relationen arbetsgivare och arbetstagare. ”Någon” behöver inte vara en fysisk person, utan är nog oftast en juridisk person i form av ett bolag eller annan association på arbetsgivarsidan eller en arbetstagarorganisation på arbetstagarsidan (arbetstagarorganisationer antas regelmässigt vara ideella organisationer). Den som vidtar åtgärden förutsätts dock ha rättskapacitet eftersom följden av ett brott mot föreningsrätten kan sanktioneras med skadestånd (54–55 §§ medbestämmandelagen). Om den som påstås ha kränkt föreningsrätten, som arbetsgivaren i AD 2013 nr 5, är en juridisk person räcker det inte med att den föreningsrättshandlingen utförs på arbetsgivarens sida. Den måste också enligt Arbetsdomstolens rättspraxis kunna knytas till arbetsgivaren i snävare mening, till en behörig företrädare. Också underlåtenhet att göra något som skulle ha gjorts kan vara en åtgärd. I AD 2013 nr 5 är den föreningsrättskränkande åtgärden att medlemmar i en viss fackförening uteslutits från en middag bekostad av arbetsgivaren, på grund av sitt medlemskap – enligt ett anslag som sattes upp i arbetsgivarens lokaler gällde inbjudan ”för alla utom de som är fackligt anslutna till if metall.” Såsom inbjudan var formulerad var den alltså utan tvekan direkt diskriminerande mot medlemmarna i en viss fackförening. Problemet var att det inte var bolagets VD eller någon annan behörig företrädare som satt upp inbjudan, utan en anställd som tagit initiativ till middagen. Det hade varit en tid med mycket arbete, och såvitt den anställde kände till hade IF Metalls medlemmar redan bjudits på middag av sin fackförening, bekostad av ett skadestånd föreningen fått av bolaget. Det föreningsrättskränkande uppsåtet finns alltså hos den anställde. Facket gjorde gällande att den anställde var arbetsledare, vilket bolaget förnekade. Arbetsdomstolen fann inte visat att den anställdes ställning var sådan att han kunde anses företräda bolaget. Att den anställdes uppfattningar om hur produktionen skulle bedrivas

[17] Se Malmberg, Jonas, Björknäs, Hanna, Eriksson, Kurt, Hansson, Mikael, Herzfeld Olsson, Petra och Larsson, Tommy, Medbestämmandelagen. En kommentar Del I, Norstedts Juridik 2018 s. 72 ff. med hänvisningar.

[18] ”Någon”-frågan finns också på den utsattes sida, och i diskrimineringslagen. Se Hellborg, Diskrimineringsansvar s. 249 ff.

till följd av dennes erfarenhet och kunskap spelade ingen roll, utan det avgörande är att ha inte var arbetsledare i formell mening. Eftersom de som hade sådan formell ställning inte kände till tilltaget att utesluta de fackliga medlemmarna från inbjudan kunde inte heller den åtgärd som uteslutandet innebar tillräknas arbetsgivaren. Arbetsgivaren dömdes ändå att betala skadestånd, men på grund av åtgärden att betala för middagen sedan det blivit känt för bolaget (det vill säga för dess behöriga företräde) att medlemmarna i fackföreningen uteslutits.

Skulden måste kunna kopplas till arbetsgivarens person, även om det är en juridisk person. Den kopplingen är formell, i att en åtgärd skall vidtas av någon på ena sidan läggs alltså att denne någon rättsligt skall knytas till sidans rättshandlingsförmåga på ett visst sätt. Skulden måste vara arbetsgivarens. Vore åtgärden att sortera ut de fackliga medlemmarna istället utförd av ett tekniskt system som arbetsgivaren infört kan det knappast tänkas att arbetsgivaren på samma sätt skulle kunna freda sig med en hänvisning till att den föreningsrättskränkande handlingen inte utförts av denne, ännu mindre att den inte utförts på dennes sida.

C-81/12 Asociatia Accept

Den svenska tillämpningen av ”någon på arbetsgivar- eller arbetstagarsidan” är restriktiv på så sätt att de möjligheter att utdöma ansvar som ges av ordalydelsen inte utnyttjas, utan de begränsas av den formella kopplingen till arbetsgivarens identitet. EU-domstolen har givit uttryck för en annan inställning vid tillämpning av det EU-rättsliga likabehandlingsdirektivet som förbjuder diskriminering i arbetslivet diskriminering i arbetslivet på grund av religion eller övertygelse, funktionshinder, ålder eller sexuell läggning.[19] Målet C-81/12 *Asociatia Accept* rörde situationen att en person som var delägare i en professionell fotbollsklubb och som av media och allmänhet ansågs vara företrädare för klubben offentligt uttalade att klubben inte tänkte anställa en viss fotbollsspelare som påstods vara homosexuell. I uttalandena framgick att det inte spelade någon roll om spelaren ifråga var homosexuell eller inte, utan att det som rapporterats i media om dennes sexuella läggning räckte för att klubben skulle vägra anställa honom. Att uttalandena i och för sig var diskriminerande var utom fråga, och att de skulle omfattas av reglerna om de gjorts av

[19] Rådets direktiv 2000/78/EG av den 27 november 2000 om inrättande av en allmän ram för likabehandling i arbetslivet.

en behörig företrädare för klubben. I domen hänvisas till mål C-54/07 *Feryn*, i vilket en företrädare för en arbetsgivare i media uttalat att hans företag inte kunde anställa utländska arbetstagare eftersom företaget då inte skulle få några uppdrag (företaget installerade säkerhetsportar). I *Accept* ställdes frågan om uttalande av någon som inte är behörig företrädare omfattas av regleringen. Efter att ha konstaterat den skillnaden mellan de båda målen (p. 46) hänförde sig EU-domstolen till bevisbörderegeln i diskrimineringsmål (som är densamma som i mål om föreningsrättskränkning enligt medbestämmandelagen). Enligt den regeln är det den som anser sig diskriminerad som har att visa omständigheter som ger anledning att anta att diskriminering förekommit (med EU-domstolens formulering), medan den som påstås ha diskriminerat har att visa att diskriminering inte förekommit. Istället för att kräva en formell koppling mellan den som gör uttalandet EU-domstolen att arbetsgivaren inte kan freda sig med avsaknaden av en sådan koppling: ”*Endast* [min kursiv] den omständigheten att sådana uttalanden som de som är i fråga i det nationella målet inte direkt görs av en viss svarandepart utgör inte med nödvändighet hinder för att det, med avseende på den parten, kan anses ha visats att det föreligger ’fakta som ger anledning att anta att det har förekommit … diskriminering’ i den mening som avses i artikel 10.1 i direktivet” (p. 48) och vidare ”[e]n arbetsgivare som är svarandepart kan således inte motbevisa att det föreligger fakta som ger anledning att anta att arbetsgivarens anställningspolicy är diskriminerande *endast* [min kursiv] genom att göra gällande att de uttalanden som suggererar att det föreligger en homofobisk anställningspolicy har gjorts av en person som, även om vederbörande påstår sig och förefaller ha en betydande ställning i arbetsgivarens verksamhet, juridiskt sett inte har firmateckningsrätt vad gäller anställningar” (p. 49). Jämfört med de ovan diskuterade svenska målen är alltså utgångspunkten den omvända, istället för att som Arbetsdomstolen i de ovan redovisade fallen kräva en formell koppling mellan den som utför den diskriminerande handlingen och den som är ansvarig för diskrimineringen (arbetsgivaren) innebär EU-domstolens tillämpning att arbetsgivaren inte kan freda sig från ansvar genom att hänvisa enbart till frånvaron av en sådan koppling. Bevisningen till stöd för att diskriminering inte förekommit skulle istället kunna vara att den anställningspolicy som faktiskt följs helt saknar samband med diskrimineringsgrunden eller att den klubb som kopplas samman med de diskriminerande uttalandena i media tar avstånd från dem (p. 56 och p. 58).

Det ändamål som skall tillgodoses med tillämpningen är att diskrimineringsförbuden skall ges verkan, vilket skulle kunna äventyras om dess tillämpning skulle begränsas av bundenhet till en formell kontraktsrelation – förbuden, och därmed arbetsgivarens ansvar, kommer alltså att tillämpas *extensivt*.[20] I den svenska tillämpningen tillgodoses inte samma ändamål, istället tycks den följa en logik som leder till en restriktiv tillämpning, att ansvar inte åläggs arbetsgivare när diskriminerande handlingar inte kan tillskrivas behöriga företrädare. Ingen av ingångarna gör det dock svårare att bortse från diskriminerande handlingar som följer av en automatiserad behandling. Det kan noteras, även om det inte spelar någon roll för de rättsliga implikationerna, att automatiserade processer för rekrytering inte är immuna mot att ta fördomsfulla eller osakliga hänsyn – tvärtom kan det visa sig, eftersom de riskerar att okritiskt reproducera osaklig särbehandling som genom tidigare särbehandling byggt in i det material som processen utgår ifrån.[21] Det kan knappast vara i linje med det effektiva genomförandet av diskrimineringsförbuden i EU-rätten att låta arbetsgivare freda sig med hänvisning till automatiserade processer.

4 Avslutning – landning i ett gammalt svar

Rädsla för förlorade arbetstillfällen är ett politiskt problem, som rätten själv är både indifferent i förhållande till och oförmögen att själv lösa. Däremot kan rätten prioritera det konservativa eller omställningen. I sig själv tycks inte de regler som omger ”arbetsgivarbegreppet” och därmed arbetsgivarens identitet utesluta andra rättssubjekt än de invanda fysiska personerna och associationerna i och för sig – om det är meningsfullt att utkräva något ansvar är en annan fråga (se strax nedan). Frågan om ansvar berörs i de tre rättsfallen som diskuteras, där en skillnad märks mellan de svenska rättsfallen från Arbetsdomstolen och fallet från EU-domstolen. I de förstnämnda (AD 2007 nr 45 och AD 2013 nr 5) kommer arbets-

[20] Hellborg, Diskrimineringsersättning s. 152 f. Hellborg drar slutsatsen att den svenska rätten inte är oförenlig med EU-rätten, eftersom det finns möjlighet att beakta EU-rätten genom en konform tolkning av 2 kap. 1 § diskrimineringslagen. Det innebär dock inte att den arbetsrättsliga förståelsen av ”arbetsgivare” och dess koppling till kontraktsrelationen med arbetstagaren inte förtjänar kritik.

[21] Se Kullmann, Miriam, Discriminating Job Applicants Through Algorithmic Decision-Making (January 1, 2019). Available at SSRN: https://ssrn.com/abstract=3373533 or http://dx.doi.org/10.2139/ssrn.3373533, s. 5 och s. 13.

givaren undan ansvar genom att andra personer utför diskriminerande handlingar. Såsom Arbetsdomstolen tillämpar diskrimineringslagen respektive reglerna om föreningsrätt förutsätts en formell koppling, eller att arbetsgivaren kan göras ansvarig enligt regler om principalansvar, principer om *culpa in eligendo, vel instruendo, vel inspiciendo* eller något annat (inget sådant diskuterades i fallen, men här handlar det om reglernas inneboende möjligheter). Det är, mot bakgrund av domen i *C-81/12 Asociatia Accept*, möjligt att rättsläget ändrats. Oavsett tycks ansvaret vara mer omfattande om arbetsgivaren använder sig av AI, eller något annat icke-mänskligt. Det finns då ingen annan att skylla på, och reglerna pekar ut ”arbetsgivaren” som ansvarig. Vare sig arbete eller arbetsledning förutsätter dock för sin existens intelligens, vare sig artificiell eller annan. De civilrättsliga reglerna syftar till ansvarsfördelning, men det kan fördelas hur som helst så länge som subjektet kan ha en förmögenhet. Regler som fördelar ekonomiskt ansvar bryr sig dock sällan om ifall den som pekas ut kan betala i och för sig, utan den ekonomiska risken stannar någonstans. De civilrättsliga reglernas grundar sig på en funktionalistisk ideologi.[22] Så länge det handlar om att ge rättigheter möter det inga större hinder att ge djur, floder eller datorer status av rättssubjekt,[23] men att ge ansvar är en annan fråga. Det är nog också där hindret för att betrakta AI som en arbetsgivare ligger. Reglerna förutsätter inte bara att någon kan åläggas ekonomiskt ansvar, utan att någon kan åläggas *skuld*.[24] Djur, floder, datorer och AI kan inte bära skuld. Skuld förutsätter ett moraliskt ansvar, och därmed mänsklighet.[25] Det är inte *artificiell intelligens* som är problemet, utan att *artificiell skuld* inte är möjlig att tillskriva. Skiftet kommer om vi tillskriver AI moral – då är den inte artificiell längre, och problemet upplöses. Till dess lär vi oss mer om rätten och dess gränser genom att läsa en gammal bok som Herman Melvilles *Billy Budd, Sailor* än alla samtida analyser av AI och rätten (inklusive denna, varför det boktipset kommer lite sent).[26]

[22] Almkvist, Gustaf, Förmögenhetsbrott och förmögenhetsrätt. Om straffansvaret i 8 och 10 kap. brottsbalken och dess förhållande till civilrätten, Iustus 2021 kap. 3.3. och s. 146 f.

[23] Fridström-Montoya, Homo juridicus s. 37 f.

[24] Den rättsrealistiska ideologin som ligger bakom den civilrättsliga samtida förståelsen av världen döljer det, se Stenlund, Rättighetsargumentet i skadeståndsrätten s. 105 f.

[25] Fridström-Montoya, Homo juridicus s. 74 och s. 80.

[26] För tips på fler, och mer samtida, böcker se Anni Carlssons bidrag i denna volym.

Vladimir Bastidas Venegas

Personalized Pricing, Discrimination and EU Competition Law

1 Introduction

By now, most of us have become used to receive offers and advertising when surfing the web. As soon as we enter into webpages with ads, it is not uncommon that we are offered products that are based on searches or online shopping that we have recently done. Websites and online shops which we visit and in which we are members also keep offering us ads and "special offers" based on our searches and our previous purchase patterns, referred to as *personalized offers* in this paper. While I am certain that many feel a certain annoyance of how quickly companies are tracing and spreading the data that we leave behind, it is also likely that many take advantage of those special offers that are made to them. We benefit from receiving discounts on products that we purchase frequently and from the "good deals" we make on products that we ordinarily cannot afford to buy. Leaving aside the issue of how companies handle our data, which is not the topic of this paper, the question could be asked whether the offers that have been adapted to a person's specific needs and demand could somehow be harmful. In particular, it could be discussed whether personalized offers constitute a form of illegal differential treatment between different customers. Such an issue seems *prima facie* to concern rules on consumer protection and anti-discrimination.[1] However, differ-

[1] F Zuiderveen Borgesius, 'Price Discrimination, Algorithmic Decision Making, and European Non-Discrimination Law' (2020), 31 European Business Law Review 401 (Zuiderveen Borgesius 2020); T de Graaf, 'Consequences of nullifying an Agreement on Account of Personalised Pricing' (2019), European Consumer and Market Law (de Graaf 2019).

ential treatment of customers could also potentially be captured by EU Competition Law when a supplier has market power. That is topic of this paper.

Although it is dangerous to make too general statements about the competition rules, a simple description of the scope of competition law is that it mainly deals with *market power*. With market power it is meant that an undertaking to a certain extent may determine market conditions, such as price and output. By contrast, under conditions of effective competition undertakings would normally have to adapt themselves to market conditions in order to survive on the market, being so-called price takers. A dominant undertaking may thus exploit its market power to the detriment of competition, competitors and ultimately the consumers. EU competition law, in particular through Article 102 Treaty of the Functioning of the European Union (TFEU), sets limits for the exercise of market power. While the pro and cons of price discrimination has always been a debated issue, it is in particular in the presence of market power that the potential negative effects of price discrimination may emerge. In cases of market power, customers are normally also not in a position to negotiate the price or other conditions. Normally, in such cases the majority of the customers will be dependent to some extent on the company with market power. In addition, if the product has the character of a must-have product, the dependency of the customers will be greater. For such cases, competition law *may* offer remedies. However, a problem is that it is debatable whether the rules are well designed to handle undertaking's behavior directly directed towards natural persons (hereinafter designated as *end-consumers*), and whether the previous case law and administrative practice give sufficient support and guidance for applying the competition rules to such situations. It may seem ironic that competition law, which is regularly characterized as promoting consumer welfare, is not self-evidently applied to all transactions directly involving dominant undertakings and end-consumers, and which may have a negative effect upon the latter. However, the main thrust of competition law is to regulate companies' behavior on the market towards customers and competitors, which supposedly will *indirectly* benefit end-consumers by promoting more competitive and efficient markets.

Accordingly, this article explores the possibility to apply Article 102 TFEU to personalized prices. Section 2 explains the meaning of personalized prices and their effect on economic welfare. Section 3 describes, in general, Article 102 TFEU and some its elements relevant for this paper. In Section 4 the possibility to capture personalized offers as illegal price discrimination under Article 102 TFEU is discussed. Finally, in section 5, it is concluded that it is problematic to assess personalized pricing under Article 102 TFEU, as the available tests may either be over or underinclusive. Arguably, it would be a better option to *primarily* regulate such practices with other bodies of rules that are focused on directly protecting end consumers.

2 Data, Algorithms and Personalized Offers

The business model of many of today's tech companies, relies on the collection of data. Online platforms, such search engines, online sales apps, or just any application or website, collect data of individuals regarding searches, location data, purchases, age, gender etc. With the development of new algorithms and data mining, such data is processed in order to create profiles of individuals that are used by online platforms to predict those individuals' behavior. Accordingly, online platforms can use such data to make individually targeted advertising and sales offers with a price only offered to the individual in question, so called personalized pricing.[2] Personalized pricing is said to consist of two elements. The first element is the practice of discriminating prices to different consumers. The second element is that the offers (normally a price) are adapted based on information of the consumer's personal characteristics and conduct (so-called targeting or profiling). The second element is normally based on the data that has been collected about the consumer.[3]

On the surface, the effects of personalized offers are ambiguous. It may be doubtful that many consumers would complain that offers are personalized as the design of such offers is adapted to our individualized possibility to pay. It seems also doubtful that many consumers would object to targeted advertising of goods and services that they are interested in. On the contrary, it may be suspected that many of us would see the benefits of such offers. For instance, a would-be customer that has visited

[2] M Botta and K Wiedemann, 'To discriminate or not to discriminate? Personalised pricing in online markets as exploitative abuse of dominance' (2020), 50 European Journal of Law and Economics 381 (Botta & Wiedemann 2020), 382; C Townley, E Morrison and K Yeung, 'Big Data and Personalized Price Discrimination in EU Competition Law' (2017), 36 Yearbook of European Law 683 (Townley et al. 2017), 684–685.

[3] OECD, 'Personalised pricing in the digital era' (2018), Background Note, DAF/COMP (2018)13 (OECD 2018), 8.

a website looking for a particular product and who has chosen not to purchase, would hardly react negatively if he/she receives an offer a few days later regarding the same product for a "discount" price. Similarly, a consumer may not object to continuously receiving new offers about such goods in the future (at least, until the demand for that good/service has been satisfied). On the other hand, it seems also that individuals may object to that "special offers" are being made to other customers, but not to them.[4] So, in other words, while we may like personalized offers when we benefit, we may actually dislike companies making such offers when we do not get the same treatment as other customers.

From an economic perspective, the discussion concerns whether a company could use personalized prices to engage in price discrimination that enhances welfare and in particular consumer welfare. According to mainstream economic theory, three conditions need to be met for price discrimination to be feasible. Firstly, there must be market power. Secondly, it must be possible to segment the market according to consumer's willingness to pay, which also presupposes the capacity to measure consumers' willingness to pay. Thirdly, it must be possible to prevent arbitrage (which is the possibility for another company to exploit price differences between customers and customer groups by buying goods/services from the low-price paying customer group and selling to the high-price paying customer group). The exact degree of market power necessary to be able to maintain a price discrimination scheme has been subject of discussion, as well as whether market power is necessary at all.[5] Although this is an interesting discussion, it is not necessary to elaborate further on this issue in this paper as Article 102 TFEU does not apply to situation when there is no or a very limited form of market power.

According to economic theory, there are three types of price discrimination.[6] Firstly, there is *perfect price discrimination*, also called *first-degree price discrimination*. This means that the supplier would be able to profile and target the willingness to pay of *each* individual customer when such a price is above the marginal cost. In such a situation, the sup-

[4] Botta & Wiedemann 2020, 388.

[5] See for instance L Henriksson, *Konkurrensrättsöverträdelser – Ekonomisk analys i den juridiska processen* (Norstedts Juridik, Stockholm, 2013) (Henriksson 2013), 174; Townsend et al. 2017, 698, cross-referencing to M Levine, 'Price discrimination without market power' (2002), 19 Yale Journal on Regulation, 1 (Levine 2002).

[6] R O'Donoghue and J Padilla, *The Law and Economics of Article 102 TFEU* (2nd ed., Hart Publishing, Oxford, 2013). (O'Donoghue & Padilla 2013), 782–783.

plier would be able to satisfy demand that could not be satisfied with a uniform market price. The effects of perfect price discrimination are mostly seen as beneficial for welfare. However, while aggregate efficiency may increase, overall consumer welfare may decrease when compared to a uniform market price. Normally, perfect price discrimination is not feasible in real markets. Secondly, *second-degree price discrimination* concerns the situation when suppliers set different prices depending on the quantity purchased by customers. This would correspond to a quantity discount given to customers. The differential pricing is not based on the identity of the purchaser but only the quantity bought. As this type of price discrimination is based on differences in costs, it is generally seen as welfare-enhancing within economic theory. Thirdly, *third-degree price discrimination* refers to the situation when different customer groups are offered different prices depending on their willingness to pay. Normally, a supplier would segment the market depending on different group of purchaser's demand elasticity. Demand elasticity, if explained in a simple and non-economic manner, refers to the sensitivity of customer groups to changes in price (or other economic factors) which affect their demand. If price increases and the demand of a customer group diminishes slightly, or not at all, there is low demand elasticity for that particular group of customers. In a third-degree price discrimination scheme, a higher price is charged to customer groups with low demand elasticity, while a lower price is offered to groups with high demand elasticity. Examples of such price discrimination may e.g. be lower prices offered for a particular service to students or elderly people. Third degree price discrimination may be welfare enhancing, as it would increase output for customers groups with a high demand elasticity. On the other hand, it could also result in higher prices and reduced welfare for customers with low elasticity.

While the account, so far, has addressed the classical view on price discrimination, which are based on economic models involving a monopolist or undertakings that are close to a monopolist, the assessment of price discrimination in imperfect competitive markets becomes more complicated. Arguably, whether the differential pricing is based upon that certain customers are willing to pay a higher price because they are either loyal to a brand or because of high search costs may have an impact on the welfare effects.[7] However, it is questionable whether this would have

[7] Townsend et al. 2017, 691–694.

an impact on the assessment of cases under Article 102 TFEU, as the provision requires that one undertaking dominates the relevant market.

It follows that the welfare effects of price discrimination are ambiguous.[8] Arguably, the only type of price discrimination that seems to have positive effects with more certainty is second-degree price discrimination. However, in real markets, not even such discrimination is unambiguously beneficial for the market, which is shown by the practice of antitrust authorities regarding rebate systems.[9] As regards the potential positive effects of price discrimination of the first or third degree, they require a certain degree of information (about the willingness to pay) and that customers could not engage in arbitrage, meaning the possibility to offer goods/services purchased to a lower price to customers that normally pay a higher price.[10] Costs for acquiring the information of the willingness to pay or to prevent arbitrage may results in overall negative welfare effects. It is therefore unclear and case specific whether price discrimination will result in positive welfare effects.

As regards discrimination through personalized pricing, the potential benefits would follow primarily from the possibility to engage in perfect price discrimination. As stated above, it is however generally argued that perfect price discrimination is not possible in real markets. The current literature therefore explores the benefits of personalized pricing in the form of third-degree price discrimination. It seems obvious that the gathering of data has made it possible for companies to acquire enough data in order to better approximate the willingness to pay in a manner that has not been possible before. To which extent this is actually costless is so far not discussed in the current scholarly writing in antitrust law. Speculating, it does not seem obvious that the development of algorithms and data collecting technology as well as their application would be negligeable. If that is correct, it may decrease the potential benefits from personalized pricing. In addition, it seems also uncertain to what extent customers, with time, would find it profitable to engage in arbitrage, considering that they (nowadays) also benefit from easy access to online sales channels for "second-hand" products.

[8] O'Donoghue & Padilla 2013, 782.

[9] See e.g. Case C-95/04 P British Airways plc v. Commission of the European Communities (EU:C:2007:166) (*British Airways*).

[10] M Borreau, A de Streel and I Graef, 'Big Data and Competition Policy: Market Power, Personalised pricing and Advertising' (2017), CERRE project report (Borreau et al. 2017), 39.

All in all, the welfare effects of personalized pricing seem to be ambiguous. While algorithms and data collection technology permit companies to easier measure customer's willingness to pay and engage in price discrimination, it is not certain that this could be done costless. Considering that the normal outcome for consumers of price discrimination is not entirely positive, additional costs for collecting data and preventing arbitrage may be problematic for endorsing a positive view on personalized pricing from an economic perspective.

3 EU Competition Law, Abuse of Dominance and End-Consumers

As noted above, the welfare effects of price discrimination and personalized pricing are ambiguous according to economic theory. That factor, as such, could be the basis of an argument against EU competition law intervention in price discrimination cases. This has however not hindered the Commission to have a somewhat harsh stance towards price discrimination, in particular by dominant firms. It is not uncommon for instance that the Commission introduces requirements through soft law of non-discrimination by dominant suppliers or suppliers with some market power when dealing with more specific situations.[11] Accordingly, it could be argued that the Commission seems to have made a policy choice that price discrimination is seen as something problematic and which preliminary should be seen as anti-competitive in the presence of market power.

If such a position is taken for granted, it is important to note that it is still not self-evident that the application of EU Competition Law could cover actions directly discriminating consumers such as personalized pricing. Although the Court has consistently made a general statement that Article 102 TFEU cover actions that directly harm consumers,[12] in principle, all cases that come under scrutiny under competition law concern actions that undertakings have taken against other companies,

[11] See e.g. Communication from the Commission, Guidelines on the application of Article 101 of the Treaty on the Functioning of the European Union to technology transfer agreements, EUT [2014] C 89/3, para. 261.

[12] Case 6/72 Europemballage Corporation and Continental Can Company Inc. v. Commission of the European Communities, EU:C:1973:22 (*Continental Can*), para. 26; Case C-52/09 Konkurrensverket v. TeliaSonera Sverige AB, EU:C:2011:83 (*TeliaSonera*).

either competitors, or customers that purchase input goods or services. The case-law is also unclear on whether the more specific prohibition of anti-competitive price discrimination could be stretched to transactions involving consumers. Although there are a number of academics that find support for the application of EU Competition Law directly to supplier-consumer relations, including price discrimination schemes, this is still an unclear issue that needs to be briefly discussed.[13]

Accordingly, in this section, the first subsection introduces Article 102 TFEU briefly, as this publication is partly directed towards non-competition lawyers. The second subsection addresses the original aims of the prohibition of discrimination under Article 102 TFEU (or then Article 86 EEC). Thirdly, the final subsection addresses the applicability of Article 102 TFEU to supplier-consumer transactions.

3.1 An overview of Article 102 TFEU

Article 102 TFEU prohibits abuse by an undertaking in a dominant position. Importantly, dominance requires definition of the relevant markets.[14] This is a complex issue, which requires an assessment of substitutability between different goods and services. The purpose of defining markets is to determine which suppliers of goods and services that exert an immediate competitive pressure on the company that is subject to an investigation under Article 102 TFEU.[15] Suppliers of goods/services that are substitutable exert such a competitive pressure. Once the relevant market has been determined, it needs to be assessed whether the company under investigation is dominant on that market. An assessment is made of the company's market share, the competitive pressure by existing competitors in the relevant market, potential competition that exerts or may exert pressure on the company, as well as whether the company meets customers with market power (buyer power). If an overall assessment indicates that the company can behave to a certain extent inde-

[13] See e.g. P Akman, 'To abuse, or not to abuse: discrimination between consumers' (2007), 32 European Law Review 492 (Akman 2007); I Graef, 'Consumer Sovereignty and competition law: from personalization to diversity', 58 Common Market Law Review 471 (Graef 2021).

[14] Case 27/76 United Brands Company and United Brands Continentaal BV v. Commission of the European Communities, EU:C:1978:22 (*United Brands*), para. 10.

[15] Commission Notice on the definition of relevant market for the purposes of Community competition law OJ [1997] C 372/5 (*Relevant Market Notice*), para. 13.

pendently from competitors, customers and ultimately consumers, the company is classified as a dominant undertaking.[16] The power to behave independently could for instance be the possibility to determine price in the market, without having customers and consumers migrating to competitors.

Once the company has been found to be dominant, it is also necessary to find an abuse, as dominance, as such, is not prohibited under Article 102 TFEU.[17] It is only abusive behavior by dominant undertakings that is prohibited, as such behavior constitutes the exploitation of market power to the detriment of competitors, customers and the market. It must be noted that there are no general criteria that are common to all type of abuses under Article 102 TFEU. Abuses may be divided up in three categories: exploitative abuse, exclusionary abuse and abuses that harm the single market.[18] With exploitative abuses it is meant that the dominant undertaking abuses its market power in relation to customers to gain advantages that it could not get under normal market conditions.[19] Exclusionary abuses result in the exclusion of companies from the relevant market(s) which diminishes the competitive pressure.[20] This may permit the dominant undertaking, at a later stage, to exploit its market power towards customers and consumers. The category of abuses that harm the single market refers to practices that specifically hinder trade between Member States.[21]

Article 102 TFEU also provides a non-exhaustive list of abuses.[22] Article 102(a) TFEU concerns the imposition of unfair prices and unfair trading conditions. Article 102(b) TFEU addresses the limitation of production, markets and technical development to the prejudice of consum-

[16] *United Brands*, para. 65.

[17] See e.g. Case C-209/10 Post Danmark A/S v. Konkurrencerådet, EU:C:2012:172 (*Post Danmark I*), para. 21.

[18] R Whish and D Bailey, *Competition Law* (9th ed., Oxford University Press, 2018) (Whish & Bailey 2018), 215–6.

[19] R O'Donoghue and J Padilla, *The Law and Economics of Article 102 TFEU* (2nd ed., Hart Publishing, Oxford, 2013) (*O'Donoghue & Padilla*), 214; J Temple Lang, 'Monopolisation and the definition of "abuse" of a dominant position under Article 86 EEC Treaty' (1979), 16 Common Market Law 345 (Temple Lang 1979), 345.

[20] O'Donoghue & Padilla 2013, 215.

[21] See e.g. joined cases C-468/06 to C-478/06 Sot. Lélos kai Sia EE and Others v. GlaxoSmithKline AEVE Farmakeftikon Proïonton, formerly Glaxowellcome AEVE, EU:C:2008:504 (*Lélos kai*).

[22] *Continental Can*, para. 26.

ers. Article 102(c) TFEU targets discrimination of certain trading partners when compared to equivalent transactions entered by the dominant undertaking, and which puts customers at a competitive disadvantage. Article 102(d) concerns the imposition on customers of the additional obligations unconnected to the main transaction, or so-called tying.

Although Article 102 TFEU does not include an exemption, unlike the provision in Article 101(1) TFEU on anticompetitive collusion, the Court and the Commission have developed the doctrines of objective justification and efficiency defense.[23] Strictly speaking, the possibility of justification is not an outright exemption but constitutes a part of the determination of abuse. However, the way that the Commission structures the assessment of efficiencies does not really differ from the assessment under Article 101(3) TFEU.[24]

3.2 The origins of Article 102 TFEU and discrimination

While it may seem to be redundant to discuss what the original intent of the founding fathers was with the provision in Article 86 EEC (which today is Article 102 TFEU), it is not uncommon that commentators have tried to interpret such an original intent to draw inferences for the interpretation of the provision today.

Importantly, the Spaak Report of 1956 explained the problem with "monopolies" and dominant undertakings.[25] In particular, monopolies and dominant undertakings could obstruct the benefits of dismantling barriers to trade which would follow from other rules in the Treaties. While the elimination of trade barriers would make it harder, if not impossible, for companies to engage in price discrimination (in particular price dumping) between different Member States, a dominant undertaking or a cartel would still have the possibility to engage in such price discrimination.[26] In addition, agreements between undertakings sharing markets would reestablish the division between (geographical) markets and could result in the limitation of technical progress. Moreover, the

[23] See e.g. Case C-209/10 Post Danmark A/S v. Konkurrencerådet (EU:C:2012:172) (*Post Danmark*), para. 41.

[24] Communication from the Commission, Guidance on the Commission's enforcement priorities in applying Article 82 of the EC Treaty to abusive exclusionary conduct by dominant undertakings OJ [2009] C 45/7 (Priority Guidelines), para. 30.

[25] Spaak Report (1956), available at https://www.cvce.eu/en.

[26] Spaak Report, 55.

domination of specific product markets would undermine the benefits of a larger market, the use of technology for mass production and the maintenance of competition.[27] The focus of the Spaak Report seems to have been foremost the use of competition rules to promote market integration and, in particular, to eliminate price discrimination.

It should be noted that as regards abuses of market power the Spaak Report only went into more depth as regards price discrimination. In the report it is stated that two specific requirements would apply to determine price discrimination.[28] Firstly, the purchaser would in practice have to submit to the supplier's conditions. This seems to be based on the notion that the supplier would be an unavoidable trading partner and thus probably a dominant undertaking. Secondly, the discrimination would result in an appreciable harm to competition between purchasers.

In this regard, it may be important to discuss the meaning of "competition" between different purchasers. Currently, the notion of competition between customers would necessarily imply that those customers are found in the same relevant market, as defined today. However, it could be questioned if the Spaak Report was necessarily based on the notion of relevant markets as we know it. It could be speculated whether the notion could be constructed broader, as purchasers of an input product or a reseller could be seen as being "in competition" by the simple reason that they would be engaged in the same sector. Arguably, such a broad notion of "in competition" could be supported by the early ground-breaking case, *United Brands*, that laid out the general framework for the application of Article 102 TFEU.[29] The Commission and the Court found price discrimination under Article 102 TFEU even though purchasers were active in different geographical markets and the competitive disadvantages could not possibly distort competition between them. The ruling in *United Brands* has been interpreted as being motivated by market integration aim and that the requirement of competitive disadvantage, in practice, was not given any real meaning. However, it could also be argued that competitive disadvantage would refer to a notion of being disfavored, which would limit the possibilities for these customers to pur-

[27] Spaak Report, 55–56.

[28] Spaak Report, 55.

[29] Case 27/76 United Brands Company and United Brands Continentaal BV v. Commission of the European Communities, EU:C:1978:22 (*United Brands*).

chase more products and the possibilities for them to resell more quantities in their respective markets.

Obviously, such a notion of competitive disadvantage rings false for competition lawyers today.[30] However, in a pure "trade context", it does not seem far-fetched that a competitive disadvantage could be construed when comparing two companies that are not active in the same geographical market. A dominant company that price discriminates customers located in different Member States can contravene the benefits of internal market, which presumedly would gravitate towards an elimination of price differences between Member States. Such price discrimination permits the dominant undertaking to enrich itself by capturing the profit margins each individual Member State can bear. Consequently, the customers' competitiveness would suffer, including the possibilities to engage in cross-border trade inside and outside of the internal market. From this perspective, the origins of the prohibition against price discrimination, seems to have been intrinsically linked to exploitative abuse and the market integration imperative.[31]

3.3 The application of Article 102 TFEU to transactions between dominant undertakings and end-consumers

It follows from Article 102(c) TFEU that price discrimination may be an abuse. The provision requires that the trading partner that is discriminated suffers a competitive disadvantage. Importantly, the provision is specific to the extent that it specifies that the aggravated party is a "trading partner". It has therefore been argued that Article 102(c) TFEU, by its wording, does not capture price discrimination practices that are directed at customers that are consumers, as such.[32] As stated above, the majority of cases under Article 102 TFEU have concerned transactions between dominant undertakings and customers that have been undertakings, or behavior directed towards other competitors (which naturally have been undertakings). This does not mean that Article 102 TFEU could not apply *at all* to transactions involving end-consumers. In fact,

[30] Se e.g. D Geradin and N Petit, 'Price Discrimination under EC Competition Law: The Need for a Case-by-Case Approach' (2005), The Global Competition Law Centre Working Papers Series 07/05 (Geradin & Petit 2005), 41.

[31] Temple Lang 1979, 346, 353 and 359.

[32] Henriksson 2013, 193.

the Court has stated that Article 102 TFEU may be applied to actions that directly injure consumers.[33] There is also a consistent view among commentators that Article 102 TFEU could apply to such transactions. However, there are only a few odd cases that actually concern behavior directed towards end-consumers. Obviously, for personalized pricing to be considered as an abuse, it is a crucial question whether Article 102 TFEU could be applied to discrimination in transactions between dominant undertakings and end-consumers. If there is no support for such an approach, it would not matter how discrimination is classified as an abuse. Accordingly, this subsection explores when and to which extent the prohibition in Article 102 TFEU applies to transactions involving end-consumers, by particularly reviewing the relevant cases under points (a), (b) and (c) of the provision, which are the points that so far have been applicable to exploitative abuses.

The preliminary question is thus to which extent Article 102 TFEU captures actions directed at consumers. As a starting point, EU Competition Law aims, amongst other things, to protect consumers or consumer welfare.[34] However, the use of the concept of "consumers" according to EU Competition Law is ambiguous as it does not refer solely to end-consumers, but includes more broadly users of a good or service, which in many cases concern undertakings acting as customers.[35] Importantly, Article 102 TFEU has mostly been used to capture exclusionary conduct, in other words, actions that would restrict competition by competitors to a dominant company. Such behavior mainly falls under Article 102(b) TFEU, the limitation of markets, although some abuses fall under Article 102(a) and (d) TFEU. As competitors are excluded, it may lead (or be presumed) that the dominant company would be enabled to increase prices or to impose unfavorable conditions on consumers.[36]

However, Article 102(a) TFEU gives support for that the provision can be applied to actions directed at consumers. The provision states that

[33] Continental Can, para. 26.

[34] T Eilmansberger, 'How to Distinguish Good from Bad Competition under Article 82 EC: in search of clearer and more coherent standards for anti-competitive abuses' (2005), 42 Common Market Law Review 129 (Eilmansberger 2005), 133–134.

[35] Commission Notice, Guidelines on the application of Article 81(3) of the Treaty, OJ [2004] C 101/97 (*Exemption Guidelines*), para. 84.

[36] N Kroes, 'Tackling Exclusionary Practices to Avoid Exploitation of Market Power: Some Preliminary Thoughts on the Policy Review of Article 82' (2005), 29 Fordham Int'l L.J. 593 (Kroes 2005), 595.

an abuse may be the imposition of unfair prices or other unfair trading conditions. In particular, the provision has been used to capture cases on excessive pricing, which concerns prices that bear no reasonable relation to the economic value of the product or service provided by the dominant undertaking. Normally, such conduct could not only be directed at customer that are undertakings, but also end-consumers. In fact, the greatest danger with dominance is that the company would engage in exploitation of market power by e.g. setting excessive prices towards customers, including end-consumers. Moreover, the situations of customers being undertakings or end-consumers are in principle the same, the only difference being that an undertaking would potentially be able to pass on excessive prices to its customers, which could include end-consumers. The case-law is however mainly concerned with the complaints made by traders that are customers of the dominant undertaking. For instance, there is a long line of case law that concerns customers to copyright collective societies paying royalties for use of music.[37] In these cases, companies have complained that the royalties have been excessive because the of the calculation method has not reflected the economic value of the service provided by the dominant undertaking. Other cases have concerned traders' complaints about excessive prices in combination with price discrimination.[38]

An early case that may give some support that the scope of Article 102 TFEU would capture transactions involving natural persons, although not end-consumers, is *BRT II*.[39] The Court held that a copyright collective society had abused its position by imposing unfair trading conditions in its contracts with two authors. Importantly, the abuse did not occur in an ordinary supplier-customer transaction. Rather the two authors constituted natural persons that had assigned their copyright, which is an "input" for the copyright collective society. Thus, the abuse concerned the "purchasing" by the dominant undertaking. It could be

[37] Case 125/78 GEMA, Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte, v. Commission of the European Communities, EU:C:1979:237 (*GEMA*); Case C-177/16 Autortiesību un komunicēšanās konsultāciju aģentūra / Latvijas Autoru apvienība v. Konkurences padome, EU:C:2017:689 (*AKKA/LA*); Case C-372/19 Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v. Weareone. World BVBA and Wecandance NV, EU:C:2020:959 (*Wecandance*).

[38] *United Brands*, paras. 234–268.

[39] Case 127/73 Belgische Radio en Televisie and société belge des auteurs, compositeurs et éditeurs v. SV SABAM and NV Fonior (EU:C:1974:25) (*BRT II*).

argued that the two authors would be in a similar position as end-consumers, in particular considering the power balance between the parties and the fact that copyright holders may be seen as unsophisticated actors like end-consumers. On the other hand, it is important to note that the two authors would likely to be defined as undertakings under EU Competition Law. The Court never dealt with this particular issue, but the fact that a copyright holder would license its rights in exchange for remuneration means that they were involved in an economic activity. Later case law also supports such a view. So even if there are similarities between an author being a natural person and end consumers, their situations are not the same. Accordingly, the case does not provide a clear example of Article 102(a) TFEU applied to a consumer transaction.

By contrast, *General Motors* concerned in part customers that would classify as end-consumers.[40] The Court found *prima facie* that the dominant undertaking had charged excessive and abusive prices. However, other circumstances would subsequently exculpate the dominant company. The case concerned prices imposed by a dominant undertaking for certificates that were necessary for imported cars to be used in a Member State. The excessive prices targeted all imports of certain cars, both imported by parallel traders as well as by natural persons. Thus, the case partly concerned the transactions between dominant undertakings and end-consumers. Nevertheless, the excessive prices would also affect traders and would in particular permit the dominant undertaking to obstruct parallel trade.[41] As stated above, the Court ultimately found that the abusive behavior had been temporary and that the company had refunded customers for the excessive prices before the Commission had taken any action. In fact, the court records seem to suggest that the company had made an "honest" mistake when charging excessive prices for the imported cars.[42] However, as the Court preliminarily classified the behavior as an abuse, the case gives an example of behavior concerning consumers transactions classified as abuse under Article 102(a) TFEU, at least when simultaneously negatively affecting traders and/or parallel trade.

[40] Case 26/75 General Motors Continental NV v. Commission of the European Communities (EU:C:1975:150) (*General Motors*).

[41] *General Motors*, para. 12.

[42] *General Motors*, paras. 19–23.

In addition, the commitment decision in *Aspen* should be noted.[43] In the case, the Commission found that the dominant company had engaged in exploitative conduct through excessive prices and through (temporary) withdrawal of medical drugs from certain member states in order to extract higher prices. The customers in this case were Member States. Member States can hardly be regarded as undertakings that would be negatively affected by excessive prices in their further "trading" with other partners. In fact, the situation between Member States and end-consumers does not differ as neither of them would be negatively affected in the capacity to trade with others. *Aspen* is however an unusual case. The importance and urgency of the case cannot be ignored, considering the costs of medical drugs for the Member States' budgets. However, the fact that there might have been an immediate and urgent interest for the Commission to intervene in *Aspen*, does not take away the fact that Member States in such a situation are comparable to end-consumers.

As regards the provision in Article 102(b) TFEU, it is generally viewed as allowing the finding of abuse when dominant undertakings engage in actions that may directly or indirectly impact end-consumers, such as the limitation of supplies and markets. Importantly, the provision explicitly states that it captures actions "to the prejudice of consumers". The schoolbook example would be that a dominant undertaking limits supply of a product that would have the same effects as an excessive price. However, in practice, the provision has mainly been applied in cases on exclusionary abuses and state actions (in conjunction with Article 106(1) TFEU) that reduce competition and that may only have an indirect negative impact on consumers.

The only example of Article 102(b) TFEU being applied to actions directly taken against consumers is the Commission's decision in *Football World Cup*.[44] In this case the company had abused its dominant position by discriminating end-consumers located outside a particular Member State hosting a sporting event. In practice, this meant a limitation of cross-border supplies of tickets and thus a limitation of markets under Article 102(b) TFEU. As previous case law on the provision only concerned exclusionary abuse, the company argued that the Commission could not apply Article 102 TFEU to the company's behavior. The com-

[43] Commission Decision, AT.40394 – Aspen, C(2021) 724 final, 10.2.2021 (*Aspen*).

[44] Commission Decision, 1998 Football World Cup, OJ [1998] L 5/55 (*Football World Cup*).

pany did not gain a competitive advantage in relation to other competitors and there were no negative effects on market structure. The Commission rejected both these arguments on the basis that the Court had already stated in *Continental Can* that behavior towards consumers could be captured by Article 102 TFEU, in particular point (b), and because discrimination on basis of nationality was contrary to principles within Union Law.[45] Importantly, the case was not litigated before the Union Courts, which means that there is no confirmation that the Commission's arguments were correct. It is interesting to note that the Commission applied Article 102(b) TFEU instead of Article 102(c) TFEU, even though the case concerned discrimination. In addition, it seems also as the Commission emphasized the cross-border trade effect of the discriminatory practice.

Turning to Article 102(c) TFEU, it has already been stated above that the provision indicates that discrimination must concern "trading partners" and result in a "competitive disadvantage", which indicates that the provision only addresses practices directed at *undertakings* that are customers to the dominant undertaking. The case law seems also to exclusively concern customers that are undertakings.

However, one case that sometimes is claimed to concern end-consumers under Article 102(c) TFEU is *Deutsche Post – International mail*.[46] The case concerned international mail delivery, where the undertaking in charge of a statutory monopoly was found to discriminate with surcharges on a group of customers that were sending mail from another Member State. One of the more important issues in the case was, in essence, whether an intermediary (which was not discriminated) between the dominant undertaking and discriminated customers would mean that the customers could not be classified as trading partners under Article 102(c) TFEU as they were not in a direct contractual relationship with the dominant undertaking. This would also be important in cases where the discriminated customers are end consumers, as the presence of an intermediary could potentially break the link between the dominant undertaking and end consumers, meaning that the latter group could never be viewed as trading partners in such cases. The Commission ar-

[45] *Continental Can*, para. 26.

[46] Akman 2007, 497–498; Graef 2021, 484; Commission Decision 2001/892/EC of 25 July 2001, Deutsche Post AG, COMP/C-1/36.915, OJ [2001] L 331/40 (*Deutsche Post – International mail*).

gued that the senders, in fact, were to be regarded as indirect trading partners that were negatively affected on their respective markets. The fact that those senders did not have a direct contractual relationship with the dominant undertaking did not hinder them from being regarded as trading partners.[47] Moreover, the company had claimed that certain customers did not suffer a competitive disadvantage. However, the Commission referred to the statement in the decision in *Football World Cup*, that Article 102 TFEU could be applied to actions that would directly prejudice consumers. Moreover, the Commission underlined that the list of abuses under Article 102 TFEU is not exhaustive. The approach could indicate that even if Article 102(c) would not capture the senders at hand because of an absence of a competitive disadvantage, an extensive interpretation of Article 102 TFEU, as a whole and without attributing the abuse to a particular point in the provision, could nevertheless capture exploitation towards customers through an intermediary. While the case could be interpreted as giving support for capturing discrimination against end consumer considering that Commission accepted the lack of direct contractual relationship, the (potential) absence of a competitive disadvantage and the willingness to go beyond the list of abuses, it must be noted that customers, *de facto*, consisted of companies involved in the sending of larger amounts of international mail through the postal services in the home state.[48] In addition, it is also clear that the case, similar to *Football World Cup*, had a cross-border trade element.

Moreover, in *BdKEP/Deutsche Post*, a case similar to *Deutsche Post – International mail*, the Commission had also found discriminatory pricing between certain major senders of mail and commercial mail preparation firms implemented through a discount system.[49] The specificities of the case are not so important as statements made by the Commission in response to certain arguments made by the parties. The defendants (Deutsche Post) had objected, *inter alia*, against the finding of abuse under Article 102(c) TFEU arguing that the senders of mail were consumers who could not be classified as trading partners under Article 102(c) TFEU. Moreover, there was no competitive relationship between the two relevant groups of customers, the major senders of mail, and the commer-

[47] *Deutsche Post – International mail*, para. 130.

[48] *Deutsche Post – International mail*, paras. 30–67.

[49] Commission Decision of 20 October 2004, AT.38745 – BdKEP/Deutsche Post AG/Germany (*BdKEP*). Only a draft of the decision is available at the Commission's website.

cial mail preparation firms that acted as intermediaries for other senders of mail. Consequently, there could not be a competitive disadvantage. The Commission responded, firstly, that major senders of mail, *included* business customers. Accordingly, the question if Article 102 TFEU could be applied to discrimination of end-consumers was irrelevant. Secondly, the Commission stated that three types of competitive disadvantage were captured by Article 102(c) TFEU. Customers could be disadvantaged in relation to the dominant undertaking itself or other customers of the dominant undertaking. In addition, a competitive disadvantage would also exist when customer's ability to compete (in which ever market) would be impaired,[50] supported by the cases on discrimination related to the single market imperative, like *United Brands*. While the Commission's decision reiterates the arguments made in *Deutsche Post – International mail*, it is interesting that the Commission dodged the question whether Article 102(c) TFEU could be applied to transactions with end consumers. Instead, the Commission relied on that undertakings were involved in the case. Moreover, the Commission clearly excluded situations involving end consumers in its list of situations where the requirement of a competitive disadvantage would be met.

Summarizing, while there are few cases under Article 102 TFEU concerning end-consumers, there is support for the application of the provision in such cases. So far, such an application has been very limited in practice. Only *General Motors* and *Aspen* (possibly) provide examples of such an application under Article 102(a) TFEU, which is also the part of the provision that has always been interpreted as being applicable to actions taken against end-consumers. With *Football World Cup*, the Commission also opened up for the possibility to apply Article 102(b) TFEU to consumer cases involving discriminatory practices resulting in harm to the internal market. By contrast, the case-law does not give support for the application of Article 102(c) TFEU to consumer cases and the Commission seems also to have dodged the issue in *BdKEP*. However, it is important to note that the list of abuses under Article 102 TFEU is not exhaustive. Considering that excessive prices towards consumers have been found to be unfair under Article 102(a) TFEU (*General Motors, Aspen*) and that discrimination by limiting supplies fall under Article 102(b) TFEU (*Football World Cup*), it does not seem too far of stretch to

[50] *BdKEP*, para. 93.

argue for that discrimination constituting the imposition of unfair conditions on end-consumers could be captured by an extensive interpretation of the provision as a whole.

4 Personalized pricing and the classification as an abuse under EU Competition Law

4.1 Personalized pricing as prohibited price discrimination under Article 102(c) TFEU

According to Article 102(c) TFEU, the abuse consists of "applying dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage", commonly referred to as a prohibition on *discrimination*. The provision has been used by the Court in cases on exploitative abuses,[51] exclusionary abuses,[52] and abuses harming the single market[53].

It follows from the wording that price discrimination consists of four elements: equivalent transactions; dissimilar conditions; trading partner; competitive disadvantage. As regards equivalent transactions, it requires a consideration of the products or services subject to the transaction as well as the conditions for the transaction.[54] It is likely that products or services are either identical or must show a certain degree of similarity.[55] Regarding potential differences in the commercial conditions between transactions, examples given in the literature are differences in the length of contract or the timing of the transaction.[56] It is mainly cost based factors that will potentially result in two transactions being classified as non-equivalent. While the assessment of equivalency may be quite a complex issue, it is

[51] Case C-179/90 Merci convenzionali porto di Genova SpA v. Siderurgica Gabrielli SpA, EU:C:1991:464 (*Merci*); Case C-525/16 MEO – Serviços de Comunicações e Multimédia SA v. Autoridade da Concorrência (EU:C:2018:270) (*MEO*).

[52] Case 85/76 Hoffmann-La Roche & Co. AG v. Commission of the European Communities (EU:C:1979:36) (*Hoffmann-La Roche*).

[53] Case 27/76 United Brands Company and United Brands Continentaal BV v. Commission of the European Communities, EU:C:1978:22 (*United Brands*).

[54] D Gerard, 'Price Discrimination under Article 82(2)(c) EC: clearing up the ambiguities' (2005), Research Paper on the modernization of Article 82 EC, College of Europe, Global competition Law Centre (Gerard 2005), 16.

[55] Henriksson 2013, 195–196.

[56] Henriksson 2013, 197–198.

of less importance for the purpose of this paper as discussed below. Arguably, personalized pricing concerns those situations where the supplier makes a distinction mainly or solely on its data on the customer's willingness to pay (see above, section 2), which is not a cost based factor or any other element that under Article 102(c) TFEU leads to two transactions being non-equivalent. Two customers to a Swedish online platform selling books, located in Sweden, would thus get different prices on the same books depending on the predictions of the customers willingness to pay made with the online seller's algorithms. In such a situation, it does not seem that the equivalent transaction criterion would be problematic for establishing price discrimination. As the willingness to pay is the determining factor for discriminating between different customers or groups customers in price discrimination of the first and third degree, such discrimination would be captured by Article 102(c) TFEU. The only type of price discrimination that could, in theory, fall outside the scope of the provision would be second degree price discrimination on the basis of the decreased cost for the dominant undertaking for purchases that reach over a certain volume. As follows from *British Airways*, this possibility is probably quite narrow. In that case, the dominant undertaking argued that there was no equivalency between those customers that have reached certain sales target with those that had failed to reach the sales targets. The Court held that two customers which had sold the same number of tickets would receive different discounts depending on whether or not they had reached individually set sales targets. Thus, for the conditions of the transaction to determine whether two transactions are to be deemed as non-equivalent, they must be based on objective cost factors that are applied consistently towards all customers.[57]

As concerns dissimilar conditions, it has been claimed that not just any difference would be sufficient to meet the requirement.[58] It is argued that the requirement is intrinsically linked to the requirement of a competitive disadvantage implying that small differences in treatment are not likely to result in a competitive disadvantage. Differences must also be put in the context of the transactions in question. However, it is uncertain whether the case law gives support for such an interpretation. As explained above, in *British Airways* the analysis focused primarily on

[57] *British Airways*, paras. 138–139.

[58] Henriksson 2013, 199–200.

the equivalence of the transactions.[59] The assessment of equivalence was established, in particular, by looking at the conditions for rebates. As the conditions were not entirely related to an objective difference in costs, as individual customers were rewarded for reaching sales targets and these differed between different customers, the difference in commission rates given to travel agents were enough to constitute dissimilar conditions. Such an approach seems to correspond with the view that the notion of dissimilar must be assessed from the perspective of the trading partner, and not of the undertaking imposing the "dissimilar" conditions.[60] On the other hand, the Court implied in *MEO* that not any differential treatment would be captured under Article 102(c) TFEU.[61] It should however be noted that the analysis of the differential treatment in the case was analyzed under the competitive disadvantage requirement.[62] Thus, it seems as the mere differential treatment, in fact, is sufficient to amount to dissimilar conditions. The question whether applied conditions truly constitute dissimilar conditions will partly fall within the assessment of equivalent transaction. And whether those dissimilar conditions are problematic or not is an issue that is ultimately dealt with under the competitive disadvantage condition. Accordingly, it does not seem likely that personalized pricing would fail to meet the requirement of dissimilar conditions.

As concerns the requirement of a trading partner, it has partly been discussed above (section 3.3). To begin with, it appears as that Article 102(c) TFEU has always been applied to transactions involving other undertakings.[63] In addition, it follows from *Deutsche Post – International*

[59] *British Airways*, paras. 133–150.

[60] Gerard 2005, 16.

[61] Case C-525/16 MEO – Serviços de Comunicações e Multimédia SA v. Autoridade da Concorrência (EU:C:2018:270) (*MEO*).

[62] MEO, para. 26.

[63] Case T-301/04 Clearstream Banking AG and Clearstream International SA v. Commission of the European Communities, EU:T:2009:317 (*Clearstream*); Case C-525/16 MEO – Serviços de Comunicações e Multimédia SA v. Autoridade da Concorrência (EU:C:2018:270) (*MEO*); Case C-95/04 P British Airways plc v. Commission of the European Communities (EU:C:2007:166) (*British Airways*); Case 27/76 United Brands Company and United Brands Continentaal BV v. Commission of the European Communities, EU:C:1978:22 (*United Brands*); Case C-179/90 Merci convenzionali porto di Genova SpA v. Siderurgica Gabrielli SpA, EU:C:1991:464 (*Merci*); Case T-83/91 Tetra Pak International SA v. Commission of the European Communities (EU:T:1994:246)

mail that a trading partner does not require a direct contractual relation with the dominant undertaking. It is sufficient that a dominant undertaking imposes conditions indirectly through an intermediary on the customer for it to be considered as a trading partner.[64] In the literature, it is commonly argued that in light of the competitive disadvantage requirement, it is not possible to stretch the wording of the provision to transactions directed towards end-consumers. However, *Akman* has argued that consumers could sometimes be in "competition" when their demand is dependent on other consumers' demand and when not every consumer could purchase a product as there are limited supplies.[65] Such an interpretation means that also an end consumer could be classified as a trading partner. Without getting into the merits of the economic models providing support for such a view, it seems far-reaching to claim that the provision intended to capture "competition" on the demand side of a market at the level of and in between end-consumers.[66] As discussed above on the Spaak Report, it seems as Article 102(c) TFEU originally targeted distortions of competition that would harm the single market. Additionally, it follows from cases such as *British Airways* and *MEO* (discussed below) that the prohibition in Article 102(c) TFEU is supposed to capture distortions of competition in upstream markets (suppliers of input product/services to the dominant undertaking) and downstream markets (customers or trading partners). Such a standpoint does not give room for end-consumers being classified as trading partners under Article 102(c) TFEU.

The requirement of a competitive disadvantage has been interpreted as meaning that the prohibition in Article 102(c) TFEU is aimed only at so-called secondary-line injury, in other words harm caused solely to the customers to the dominant undertaking.[67] As mentioned above, both the Court and the General Court have stated that Article 102(c) TFEU is aimed at capturing distortion of competition in upstream and downstream markets caused by the dominant undertakings discrimination.[68]

(*Tetra Pak II – CFI*); Case 85/76 Hoffmann-La Roche & Co. AG v. Commission of the European Communities (EU:C:1979:36) (*Hoffmann-La Roche*).

[64] *Deutsche Post – International mail*, para. 130.

[65] Akman, Pinar, 'To abuse, or not to abuse: discrimination between consumers' (2007), 32 European Law Review 492 (Akman 2007).

[66] Townsend et al. 2017, 741.

[67] Geradin & Petit 2005, 9.

[68] *MEO*, para. 24; *British Airways*, para. 143; *Clearstream*, para. 192.

This should be distinguished from first-line injury which refers to cases when the injured party is a competitor, as e.g. when a dominant undertaking discriminates customers through a rebate system with the effect of excluding competitors.

In the literature, the common view appears to be that a competitive disadvantage cannot be derived from the smallest differential treatment, but the older cases are ambiguous. In fact, what could be read out from the case law is a development where the Court has gone from a very "light" assessment of competitive disadvantage to a more detailed and stricter assessment. As discussed above, the Court found in *United Brands* that discrimination of customers in certain Member States was captured by the provision without making an analysis of the competitive relation between customers located in different Member States. A similar assessment was also done in other cases, such as *Tetra Pak II*, which also concerned price differences between different Member States.[69] As also pointed out above, the rationale behind the judgment in *United Brands* may be that the competitive relation at that time was not determined by notions such as the relevant market. Rather, the judgment implies that companies that would be involved in the same economic activity, as the resales of bananas, could on a general level be seen to be in a competitive relation with one another. In addition, the Commission's interpretation of competitive disadvantage in *BdKEP* (see above, section 3.3) indicated that it is not necessary to show a competitive disadvantage in relation to the dominant undertaking itself or customers of the dominant undertakings. It would be enough that the trading partners' competitiveness would be affected in any market when the discrimination would harm the internal market in cases such as *United Brands* and as implied by the Spaak Report. The very purpose of Article 102(c) TFEU would thus be to prohibit discrimination of customers in different Member States, as such exploitative behavior made possible by the market power of a dominant undertaking would eliminate the benefits of having one single market.

By contrast, in later case law, such as *British Airways*, which mainly dealt with an anticompetitive rebate system resulting in primarily-line injury, the Court seems to have taken a narrower view of competitive disadvantage. In the case, it was found that competition among trading partners (travel agencies) was determined by two factors, the ability to

[69] Case T-83/91 Tetra Pak International SA v. Commission of the European Communities (EU:T:1994:246) (*Tetra Pak II – CFI*), paras. 160–173.

provide suitable flight seats to a reasonable price and the travel agencies individual financial resources. The problematic parts of the rebate system led to exponential changes in revenue with a negative impact on the trading partners' financial resources and thus their competitiveness.[70] Arguably, the Court seems to have consider, at least implicitly, whether the trading partners were active in the same relevant market and how the rebate system would impact their capacity to compete. None the less, the Court seems not to have engaged in any form of a more detailed assessment of effects. However, later in *MEO*, the Court made its most far-reaching statement on the requirement of a competitive advantage.[71] The case concerned a collective society that had a monopoly on the managing of its members' rights. It applied different tariffs towards different customers and the question arose whether there was illegal price discrimination under Article 102(c) TFEU. It follows also from the facts of the case that the price difference was low when compared to the average costs of the customer, meaning that it was uncertain whether the price difference would have any appreciable impact on the customer's competitiveness. The Court held that the notion of distortion of competition under Article 102(c) TFEU encompassed "to hinder the competitive position of some of the business partners of that undertaking in relation to the others".[72] While there is no *de minimis*-threshold, a mere disadvantage because of differences in tariffs would not be sufficient. The Court also held that it is not necessary to demonstrate any *actual* effects of trading partners being disadvantaged toward its competitors. Referring to *Intel*,[73] the Court stated that an overall assessment of a competitive advantage should include "the undertaking's dominant position, the negotiating power as regards the tariffs, the conditions and arrangements for charging those tariffs, their duration and their amount, and the possible existence of a strategy aiming to exclude from the downstream market one of its trade partners which is at least as efficient as its competitors".[74] Apart from the assessment of the duration and the amount of the tariffs, it is not self-evident that the enumerated factors are relevant to determine a potential competitive disadvantage. The last part of the statement by the

[70] *British Airways*, paras. 146–148.

[71] Case C-525/16 MEO – Serviços de Comunicações e Multimédia SA v. Autoridade da Concorrência (EU:C:2018:270) (*MEO*).

[72] *MEO*, para. 25.

[73] Case C-413/14 P Intel Corp. v. European Commission (EU:C:2017:632) (*Intel*).

[74] *MEO*, para. 31.

Court seems particularly odd. Even if the dominant undertaking would pursue to eliminate one of its trade partners, it does not seem relevant that the assessment should be focused on whether the behavior has the capacity to eliminate an as-efficient-competitor. Such a standard follows from case law on exclusionary abuses, as e.g. rebate systems that may lead to primary-line injury. If the Court's statement is correctly understood, the standard for finding a competitive disadvantage would differ depending on the trading partner that is discriminated. The threshold for finding illegal discrimination would be higher when the dominant undertaking aims at excluding a trading partner than when the dominant undertaking simply wants to discriminate. Such a difference in standard does not make sense. A strategy to exclude through discrimination a trading partner that is not an as-efficient competitor in the relevant market where it operates, would still distort competition in the downstream market. Leaving aside this particular statement by the Court in *MEO*, it seems clear that the judgment makes the assessment of competitive disadvantage considerably stricter.

It follows that the case law on competitive disadvantage is not entirely consistent. Commentators have criticized the inconsistencies, in particular the differences in the assessment between cases such *United Brands/Tetra Pak II* and *British Airways/MEO*. However, the evolution in the case law could also be simply viewed as the development of stricter standard with time. Alternatively, the case law is not necessarily inconsistent, but rather expresses the protection of different interests. Cases such as *United Brands* concern the single market imperative, where the competitive disadvantage is interpreted as a broader notion (see above, section 3.2–3.3). By contrast, cases such as *British Airways* and *MEO* concern distortions of competition between traders that are located in the same relevant market, while not causing harm to the internal market, and which therefore requires a more detailed analysis.

As found above, it is doubtful whether personalized pricing directed towards end consumers could meet the requirement of "trading partner". In addition, irrespectively of which interpretation of the requirement competitive disadvantage is accepted, it follows that it cannot be easily applied to personalized pricing. End consumers cannot suffer a competitive disadvantage as they are not active as undertakings on any market. However, similar to the situation of undertakings in different Member States being discriminated, they may be exploited through price differences. Thus, potentially, an analogy could be made between these two situations when discrimination involves and is based on end consumers being located in different Member States, as in *Football World Cup*. However, personalized pricing does not concern such situations, but rather discrimination based on the consumers' willingness to pay irrespective of their location.

If personalized prices nevertheless would be found to constitute prohibited price discrimination under Article 102(c) TFEU, there is always a possibility for justification, also on efficiency grounds. As stated above, even though Article 102 TFEU does not include any exemption, the Court and the Commission have accepted objective justifications and an efficiency defense. Arguably, the latter would be primarily interesting for arguing the positive effects of price discrimination. The dominant undertaking would be required to demonstrate an economic benefit that outweighs negative effects on efficiency and consumers; the indispensability of the abusive conduct; that the abuse does not eliminate competition in the relevant market.[75] While there are cases on justification of second-degree price discrimination (exclusionary rebate systems), it seems that the Court has not yet analyzed the possibility to justify schemes regarding first and third-degree price discrimination. What may be said is that the Unions Courts have been quite restrictive in accepting justifications. In most cases, dominant undertakings have failed to establish, as a matter of evidence, that the abuse produce the claimed efficiencies.[76] What follows from above (section 2) is that the dominant undertaking would need to proof the increase in output permitted by the personalized pricing and that such increase outweighs the effects of the higher prices charged to other customers or customer groups.

All in all, it is submitted that there is no support for finding personalized pricing as an abuse under Article 102(c) TFEU in the light of the Court's case law and the Commission's previous decisions. While the conditions of equivalent transactions and dissimilar conditions could be met, it would require a far reaching reinterpretation of the conditions of trading partner and competitive disadvantage to fit with cases on personalized pricing. Would personalized pricing, *prima facie*, still be captured by Article 102(c) TFEU, under the current state of the law, there exist a possibility for the dominant undertaking to invoke the positive effects of

[75] Priority Guidelines, para. 30.

[76] Case T-219/99 British Airways plc v. Commission of the European Communities (EU:T:2003:343) (*British Airways – CFI*), paras. 290–291.

price discrimination as a efficiency defense, even though such an argument would have difficulties to succeed.

4.2 A consumer welfare approach to personalized pricing under Article 102(c) TFEU

It was argued above that the effects of price discrimination through personalized pricing on welfare and consumers are ambiguous. Commentators have therefore argued that an assessment of personalized pricing should be adapted to adequately only target price discrimination that would result in negative effects on welfare and consumers, or a so-called effects-based approach. For instance, as argued by *Townsend et al.* and *Akman*, an abuse should only be found once it is demonstrated that price discrimination would not lead to an increased output (for those customers that would otherwise not be served) in an individual case.[77] Obviously, it has already been concluded in this paper that Article 102(c) TFEU cannot be applied to price discrimination through personalized pricing (see above, section 4.1). None the less, as the discussion in the doctrine still revolves around the possibility to apply an effects-based approach under Article 102(c) TFEU to personalized pricing, it is interesting to explore whether Article 102(c) TFEU, in theory, could give room for such an effects-based assessment.

Addressing more specifically the general criteria for finding a prohibited price discrimination in Article 102(c) TFEU, it seems as they give little room for engaging in such an effects-based analysis. The problem is that there is no criterion under Article 102(c) TFEU that gives room for making any type of estimation of effects on total welfare or consumer welfare. Naturally, price discrimination could fall outside the provision completely, if transactions with two different customers that have different levels of willingness to pay are classified as not equivalent. However, such a result would be problematic, since the effects on total welfare and consumer welfare are ambiguous and case specific. A rule that completely excludes price discrimination based on different customers' willingness to pay is too lenient, while a rule that always captures price discrimination is too strict. It seems as the only type of price discrimination that could possibly by excluded through the equivalence requirement is second-degree price discrimination, which is not relevant for the discussion on

[77] Akman (2007), 511–512; Townsend et al. (2017), 738–744.

personalized pricing in this paper. Moreover, the criteria of trading partner, dissimilar conditions and competitive disadvantage have not been designed to make a balancing of effects. In particular, to read-in an assessment of effects under the criterion of competitive disadvantage seems far-reaching. The purpose of the criterion is to measure whether the economic power of the dominant undertaking is capable of doing harm to the competitive process in the upstream and downstream markets, but by measuring *harm to its trading partners* and not the aggregate effects on the market or the collective of customers.

It follows that the type of effect-assessment of price discrimination as propagated by *Akman*, *Townsend et al.*, and others, would require a dramatic re-interpretation of the wording of Article 102(c) TFEU. Alternatively, price discrimination through personalized pricing could be deemed to always fall under Article 102(c) TFEU, if for instance, the requirement of trading partner would be given a wide interpretation and the condition of competitive disadvantage would not be applied to personalized pricing, similar to the cases on harm to cross-border trade.

It may be debated whether such an outcome would be desirable. Importantly, the prohibition against discrimination has not been designed to capture every differential treatment that constitutes discrimination. It has been argued above that the prohibition against discrimination seems to have been based on mainly two rationales: protection of the integration of the single market and distortions of competition in upstream and downstream markets because of the dominant undertaking's exploitative behavior. Irrespective of whether these two rationales are viewed as legitimate or not, what follows is that the threshold for classifying discrimination as abuse is higher than mere discrimination. Accordingly, some price discrimination schemes are not captured by Article 102 TFEU, even in the presence of market power, because there is no harm to a protected interest under EU Competition Law. Unless there are policy reasons why the threshold should be lower for cases concerning discrimination of end-consumers, it is no way self-evident that price discrimination in such cases should be found abusive.

Would such an approach however be accepted, it would open up for a balancing test through the efficiency defense based on the increased output following the price discrimination scheme. The increased output directed towards a particular group of customers could, in theory, be seen as an economic benefit that may be balanced against the higher prices for other customers. Likewise, it could be argued that the differential pricing would be indispensable to cover demand from the benefitted customers. The difficulty for the dominant undertaking would be to show indispensability and that the positive effects would outweigh the negative effects. In addition, the costs for discovering the willingness to pay and hindering arbitrage would also have to be taken into account. While an efficiency defense would permit, in theory, of a proper balancing test that would determine the effects on welfare and consumers of personalized pricing, the possibilities for dominant undertaking to defend themselves with objective justifications and an efficiency defense, in practice, seems difficult.

It follows that it is unlikely that an effects-based approach could be integrated in the application of Article 102(c) TFEU.

4.3 Personalized pricing as a non-listed abuse under Article 102 TFEU

An alternative approach, as suggested by *Townsend et al.*, is to capture personalized pricing as a non-listed abuse. As stated above, the Court has been reluctant to claim that an abuse explicitly falls outside the non-exhaustive list of abuses in Article 102 TFEU. However, such an approach has the benefit that the Court is free to design a test for a specific type of abuse, unburdened by the listed examples of abuses in Article 102 TFEU and/or its previous case law.

Arguably, this is what the Commission already did, in practice, in *Football World Cup*, when the Commission viewed the discriminatory behavior as limitation of markets with prejudice to consumers under Article 102(b).[78] In *Football World Cup*, the use of Article 102(b) TFEU seems logical considering that practice purported to sell less tickets or to limit supplies to residents outside a particular Member State. However, the circumstances in *Football World Cup* are peculiar, as normally an undertaking would have little incentive to want to stop sales to a particular customer group for reasons unrelated to competition or parallel trade. Moreover, the single market aspect of the decision should also not be ignored. In response to the claim that the dominant undertaking had not gained a commercial advantage through its behavior, the Commission pointed towards that the behavior constituted discrimination on basis

[78] Commission Decision, 1998 Football World Cup, OJ [1998] L 5/55 (*Football World Cup*).

of nationality.[79] Accordingly, the behavior by the dominant undertaking was harmful to the single market as it hindered cross-border sales of tickets. Consequently, the behavior in this case did also not merely constitute discrimination between end-consumers as the finding of abuse also required the single market element.

By contrast to the approach chosen in *Football World Cup*, the main alternative is an effects-based approach. Price discrimination is abusive only insofar it does not result in an increase of output that outweighs negative effects. There are two possible approaches to design such a rule. The first alternative would be to *prima facie* classify all discrimination through personalized pricing as abusive, but opening up an analysis of the effects on consumer welfare through objective justification. Such an approach corresponds to applying Article 102(c) TFEU to personalized pricing without requiring the proof of a competitive disadvantage and keeping a wide interpretation of the notion of a trading partner. In practice, one could imagine that a competition authority/claimant would be required to show that the dominant undertaking in a consistent and systematic manner has used tracing technologies and algorithms that collects and process data to determine individual consumers willingness to pay, which is subsequently used in the design of personalized offers that result in differential pricing. It should also be noted that this approach would, in essence, have as starting point that dominant undertakings must charge a uniform price towards all end-consumers.

One objection to such an approach would be that it opens up the floodgates to challenges by consumers under EU Competition Law. However, such an effect should not be exaggerated. Firstly, personalized pricing could only be targeted when there is a dominant undertaking. Secondly, considering the complexity of these cases in the gathering and the analysis of evidence, it seems unlikely that there would be room for much private enforcement, even by larger organizations representing consumers' interests. Thirdly, it seems also unlikely that competition authorities would prioritize such cases.

However, the main objection to such an approach, is that there is no specific factor that is identified as a *prima facie* harm under EU Competition Law that triggers the need for the dominant undertaking to justify its behavior. As stated above, cases concerning Article 102(c) TFEU have partly relied on harm to the competitive process through the competi-

[79] *Football World Cup*, recital 102.

tive disadvantage requirement. Cases concerning Article 102(b) and (c) have partly relied on harm to the single market caused by limitations to cross-border trade. Cases on Article 102(a) TFEU have relied on the imposition of excessive prices enabled by the market power of the dominant undertaking. However, personalized prices could only *potentially* cause harm by limiting supplies to certain groups of end-consumers. Importantly, by contrast to *Football World Cup*, cases on personalized pricing would require the establishment of the limitations of supplies to certain groups of end-consumers because of the higher prices imposed on them. This may be empirically difficult to proof in individual cases, unless a higher price (compared to prices by other customers) is *presumed* to cause such an effect, in which case personalized prizing would always be viewed as limiting output.

For lawyers that do not work with competition law, it may seem strange that there may be some hesitation in applying competition law to cases where at least some consumers may be harmed by paying higher prices than others, particularly if competition law aims at protecting consumer welfare. However, from a competition law perspective, differential pricing at the level of individual consumers, as such, is not obviously viewed as a form of harm that would legitimize intervention against dominant undertakings. While in the literature it has been argued that competition law could consider the vulnerability of certain groups of customers to personalized offers,[80] such a notion is also unconnected to the interests protected under EU Competition Law mentioned above. Such types of considerations are arguably more connected to rules on direct consumer protection and possibly anti-discrimination law. Rules providing direct consumer protection seem to be preoccupied with the information given to consumers or the lack thereof in connection to a transaction. Moreover, they also consider how the design of a commercial message or the content of a contract may create an unjustified imbalance between the parties enabled by information asymmetries and the imbalance between the parties' bargaining power. In addition, anti-discrimination rules are focused on the protection of certain groups of individuals based on grounds such as gender, ethnicity, sexual orientation, religious belief etc. from unequal treatment. However, EU Competition Law does not target undertakings' behavior towards consumers that exploit information

[80] I Graef, 'Consumer Sovereignty and Competition Law: From Personalization to Diversity' (2021), 58 Common Market Law Review 471.

asymmetries, general differences in bargaining power or the predisposition of certain groups of individuals. It is mainly concerned with the exploitation of market power and harm which is normally attributable to the presence of market power, such as e.g. the excessive prices that a monopolist would impose in a market. Naturally, a dominant undertaking could also engage in e.g. unfair commercial practice such as misleading consumers with advertising. And obviously it could be argued that the negative consequences of such behavior may be aggravated if the majority of consumers could not choose another supplier because the dominant undertaking *de facto* offers a must-have product. Still, it could be argued, that EU Competition Law would and should not apply to such a case. The problem with misleading advertising, namely the exploitation of the information asymmetry between the dominant undertaking and the consumer, is not a problem addressed by EU Competition Law.

It appears as there are problems with identifying a particular interest under EU Competition Law that is harmed when discrimination towards end-consumers occurs in cases unrelated to the single market imperative, distortions of competition or the exploitation through excessive prices. This may give the impression that the direct protection of consumers is somewhat unprioritized under the competition rules. That is probably true, as protecting competition mainly afford benefits *indirectly* to the *collective of consumers* through lower prices and better quality which follows from competitive markets. However, it must be remembered that also Article 102(a) TFEU may afford consumers a certain degree of protection, including cases on personalized pricing. Unfair prices and conditions may be established in cases of excessive prices as demonstrated by cases such as *General Motors* and *Aspen*. The provision would also apply in cases of personalized prices when prices charged to a particular customer group of end-consumers are excessive. Moreover, in the *German Facebook Case*,[81] the German Competition Authority also applied the provision in German competition law corresponding to Article 102 TFEU to data collection which was contrary to data protection law and therefore could be viewed as unfair. While the Union courts have not confirmed that the approach in *German Facebook Case* is a viable alternative under EU

[81] Bundeskartellamt, Case B6-22/16, Facebook, see press release at https://www.bundeskartellamt.de/SharedDocs/Entscheidung/EN/Fallberichte/Missbrauchsaufsicht/2019/B6-22-16.pdf;jsessionid=A82BB51FC236DE99DEDBBF794E961517.1_cid378?__blob=publicationFile&v=4 (*German Facebook Case*).

Competition Law, it constitutes a possible way to deal cases concerning the imposition of conditions related to data and end-consumers. Considering the value of data for undertakings, arguably, an analogy could be made to the imposition of excessive prices under Article 102(a) TFEU as the dominant undertaking could be seen as *enriching* itself by extracting excessive amounts of data from individuals. The key issue then becomes how to define "excessive" in such a context. This indicates the problem with making analogies between personalized pricing and the *German Facebook Case* is that in the latter case there was a standard set by the data protection rules on what should be classified as unfair conditions. In cases on price discrimination through personalized prices, it would be required to determine under EU Competition Law what kind of price discrimination that is deemed as unfair. Such an approach only takes us back to the effects-based approach to make a determination of positive and negative effects of personalized pricing.

The second approach to personalized pricing is to require the competition authorities to make an explicit balancing exercise in the light of consumer welfare before establishing a *prima facie* abuse. Such an approach may be beneficial as it may provide a reasonable opportunity for the dominant undertaking to be successful in cases regarding output-increasing price discrimination as the burden of proof would be on the claimant/competition authorities. The current application of the objective justification/efficiency defense has arguably been too strict towards dominant undertakings because of the burden of proof. At the same time, such an approach would still provide no quarters for dominant undertakings in cases where the behavior results in none or a very small increase of output (as arguably was the situation in *Football World Cup*). Naturally, the drawback of such an approach is that the analysis may become too complex for a claimant/competition authority and that it would inconsistent with the common structure applied to other abuses and other restrictions of competition. Normally, a claimant or a competition authority is not required to engage in full scale calculation of total welfare and consumer welfare effects to prove a restriction of competition or an abuse. Not even the so-called effects-based approach as regards exclusionary abuses requires such a full calculation of welfare effects, but only that the behavior has the capability of eliminating as-efficient competitors, which in turn would permit it to engage in future exploitative abuse. It could be speculated whether a full-scale investigation of welfare effects would in effect exclude private claimants, as they would not have the resources and the investigative tools (available to competition authorities) for gathering data to prove their case. Moreover, the regular approach under the competition rules is that the claimant/competition authority identifies *prima facie* harm to a particular interest, such as market structure, the single market and/or the competitive process, while it is up for the defendant(s) to prove those efficiencies that may follow from the investigated conduct. There is a logic behind this structure in the sense that it is the defendant that is in the best place (as a matter of burden of proof) to present and explain the efficiency rationale behind its conduct.

It follows that neither of the two proposed rules above are optimal. The two alternatives risk either over- or underenforcement of Article 102 TFEU. Considering that data protection rules deal with the purposes and extent of data collection and processing, and that other rules on consumer protection deal with the information and advertising directed to consumers regarding e.g. services provided through an app, it would perhaps be a better option to regulate such aspects of transactions involving personalized pricing through such rules and not competition law.

5 Conclusions

It has been found that the welfare effects of personalized pricing are ambiguous. Personalized prices may increase output directed towards certain group of consumers that would otherwise not be served. On the other hand, such benefits may come to a cost for other consumers groups that would pay higher prices. To which extent EU Competition Law can be applied to price discrimination with negative welfare effects is also unclear. There is clear support for that Article 102 TFEU may be applied to conduct by dominant undertakings towards consumers. It is however not so clear whether Article 102 TFEU can be applied to cases on price discrimination involving personalized pricing when there is no connection to harm to the internal market. There seems not to be support for the application of Article 102(c) TFEU to such cases. Nevertheless, in theory, a wide interpretation of Article 102 TFEU, may include cases on personalized prices. The problem with such an application is that it is difficult to find a clear competitive harm in cases on personalized prices that is recognized under the current rules. If the negative effects on certain consumer groups would be deemed as competitive harm, it would require the competition authorities to engage in a full-scale assessment of welfare effect for finding an abuse. In the alternative, Article 102 TFEU could be interpreted as always requiring uniform pricing by dominant undertakings towards consumers, unless the dominant undertaking could present an efficiency defense. Arguably, as the effects of personalized pricing are ambiguous and the possibilities to present an acceptable efficiency defense are probably narrow, it could prevent behavior by dominant undertakings, which at the end of the day are not necessarily so bad for consumers.

Mattias Dahlberg

Digital Business and Tax Law: New and Global Tax Rules for Tech-Giants Using Artificial Intelligence in their Business Models

1 Introduction

Digital business and tax law is an evolving field in many aspects.[1] The growth of the digital market has made the traditional grounds for the allocation of taxing rights less effective. The objective with this article is to present and discuss new legislation on the taxation of multinational enterprises ("MNEs") agreed upon by more than 130 nation states on 8 October 2021. The legislation is so far not decided by national parliaments, and it is still uncertain whether if for example will be passed by the US Senate. The target of the new legislation is in particular large US tech-companies conducting digital business, such as Google, Apple, Facebook and Amazon. Artificial intelligence (AI) is a core element of the business models of these tech giants. In addition, MNEs in other fields of business are also covered by the new tax rules.

The agreement generally deals with the allocation of taxing rights between states. Many states fear the loss of revenue due to the old tax law concepts not being sufficient to impose an effective taxation of business income, in particular income from digital businesses activities. A large

[1] Many thanks to Assistant Professor Katja de Vries and Associate Professor Martin Berglund for valuable comments on a previous version of this article. Of course, I am solely responsible for the views expressed and any remaining mistakes.

number of states have recently decided to apply a new model for taxing digital businesses.

How does this connect to artificial intelligence? One connection is that several of the MNEs that will be taxed according to the new set of rules are using artificial intelligence in their business models.

The proposal from the OECD, which has now been agreed upon, and is discussed in this article, is called "Pillar One". "Pillar Two" of the OECD project concerns new anti-avoidance measures supporting the traditional tax concepts, that will continue to work in parallel with the concepts of Pillar One.

The concept of artificial intelligence is only briefly addressed in this article. However, it is obvious that it has had a large impact on business which the OECD labels "automated digital services". This is at the core of the tax agreement on Pillar One as elaborated by the OECD in the 2020 Blueprint report.

Other perspectives on tax law and artificial intelligence are possible. An obvious one concerns the collection of information from and about taxpayers, and which now have also reached the activities of tax advisors.[2] The EU is active in this area of law, and a number of directives have been introduced. The exchange of information between tax agencies has for long been an area of interest to the EU. Already in 1977, the European Community adopted a directive on the exchange of information. In recent years, this has developed much further. The fundamental legal act is, today, the Directive on Administrative Cooperation (DAC), which is the basic directive.[3] This directive has in turn been supplemented with revised versions of the directive on several occasions. The so-called DAC 6 contains an obligation for tax advisors to file information to the tax agency on planned, but not necessarily executed, tax planning schemes.[4]

[2] The use of AI from a tax agency perspective is discussed in Zackrisson, Marcus; Bakker Anuschka and Hagelin, Johan, AI and tax administrations: A good match, Bulletin for international taxation (IBFD), 2020, pp. 619–625, and Antón, Fernando Serrano, Artificial intelligence and tax administrations: Strategy, applications and implications, with special reference to the tax inspection procedure, World Tax Journal (IBFD), Volume 13, No. 4 (published online 27 September 2021).

[3] Council Directive 2011/16/EU of 15 February 2011 on administrative cooperation in the field of taxation and repealing Directive 77/799/EEC.

[4] Council Directive (EU) 2018/82 of 25 May 2018 amending Directive 2011/16/EU as regards mandatory automatic exchange of information in the field of taxation in relation to reportable cross-border arrangements.

This is a fundamentally new approach by the European Union, and it raises som questions in relation to fundamental rights.

Another less discussed issue is international tax planning using digital tools. In the 1990s and early 2000s such tax planning tools were much marketed towards tax practitioners. It is my impression that such marketing is less frequent today.

What is artificial intelligence? In my view it seems rather to be a concept that can be described but hardly defined. It is a computerized process, at least to a significant degree, in which creative, innovative results are acquired. The process requires the collection of information, and the application on that information of some kind of a mathematical algorithm that identifies human connections, patterns, interests and preferences, some of which may be earlier known, while others may be previously unknown. I must admit that I am a bit sceptic to the term "artificial intelligence", perhaps because it almost seems to be an oxymoron (contradiction in terms), just like "minor crisis", "only choice", or – half-jokingly – "military music". In my mind "intelligence" in its cognitive meaning is something that includes both "sense and sensibility", to use the words of the novelist Jane Austen.[5] Or is it precisely that distinction which the attribute "artificial" in "artificial intelligence" refers to: Sense ("intelligence") without sensibility ("artificial")?[6]

The algorithms that constitute a basis for artificial intelligence may, of course, express prejudices, and even spreading and strengthening them. They may ignore minorities, for example due to unrepresentative training data, as well as the – sometimes – arbitrary and unpredictable behavior of mankind.

Still, it is true that the digitalization of the world economy has largely influenced the lives of many. A simple example can illustrate the idea. A multinational enterprise ("MNE") which provides a search engine on

[5] Austen, Jane, Sense and sensibility, Penguin Classics, London, 2004 (1811).

[6] There is a legal definition in the proposed EU Artificial Intelligence Act. The definition is contained in Article 3 and reads: "'artificial intelligence' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with". See European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, COM(2021) 206 final, Brussels, 21.4.2021.

the internet identifies that someone is searching for an apartment to buy. In response, the search engine directs advertisements on new furniture and bathrooms to the apartment seeker. This is likely to be for the benefit of all parties involved. The growth of multinational tech companies in recent years bears evidence of this development. It is likely that the covid-19 pandemic has enforced the digital evolution. In addition, the major government subsidies given to businesses during the pandemic, also enforces the need for states to secure tax revenue. Already before the pandemic, many states had responded unilaterally with different levies on income from digital business activities. The United States, where many of these tech companies are resident, have responded with sharp criticism of the measures. For example, in relation to France, the US made some threats of minor retaliatory measures, such as tariffs on champagne and other French luxury products.[7] In recent years, the OECD and G20 have been working on proposals to tax digital businesses and to establish a global minimum taxation of corporations.

Today, in November 2021, there are several European states that have implemented different kinds of taxes on digital services. Other states are planning or have proposed the introduction of such taxes.[8] The idea is that a global tax system according to Pillar One will replace such national measures, which has been strongly advocated by the United States.

As previously mentioned, an agreement has now been made between more than 130 nation states and jurisdiction on the Pillar One and Pillar Two proposals. These states and jurisdictions cover more than 90 per cent of the world's GDP.[9] However, the agreement between the involved states and jurisdictions must also be enacted with national law and tax treaty law. This might be an even larger problem than reaching the October 2021 agreement.

For several years there has been an ongoing discussion in the EU on whether to adopt common legislation on the taxation of digital services

[7] Cf. *The Economist*, July 11, 2019, France's digital tax riles the White House.

[8] Among the states which unilaterally have introduced taxation on digital services are Austria, France, Hungary, Italy, Poland and the United Kingdom. See Alvarado, Mary, Digital services taxes across Europe in the midst and aftermath of the Covid-19 pandemic: A plausible option to raise tax revenue, European Taxation (IBFD), 2021, pp. 403–410, at p. 407 (with a list of current digital services taxes in Europe either proposed or decided).

[9] OECD/G20 Inclusive Framework on BEPS, Progress report July 2020–September 2021, Paris, 2021, p. 2.

and other digital business activities. In 2018 it came so far that the European Commission proposed two directives on the issue. One directive concerned a digital services tax and the other corporate taxation of a significant digital presence.[10] These directives have been subject to considerable discussion, and today the Commission has withdrawn both of them. Instead, in May 2021 the Commission presented plans for a more comprehensive approach to corporate taxation, including the taxation of digital business.[11] The Commission plans to propose a "digital levy"[12]. The Commission has also stated that it intends to propose a directive that would implement Pillar One with EU Law.[13] An agreement on Pillar Two would also require changes with EU law affecting the Member states.[14] A recent statement by the European Commissioner for economy, Paolo Gentiloni, suggests that it is only Pillar Two that requires legislation through EU directives, and that Pillar One may not require an EU Directive.[15] The probable reason why a directive may not be required for Pillar One is that it will be based on a multilateral treaty between the states and jurisdictions involved.

In the EU there is a requirement for unanimity to decide on tax matters, such as implementing Pillar One and Pillar Two through directives.[16] For some time, it was uncertain whether Ireland, Estonia and

[10] Proposal for a Council Directive on the common system of a digital services tax on revenues resulting from the provision of certain digital services, COM(2018) 148 final, Brussels, 21.3.2018, and proposal for a Council Directive laying down rules relating to the corporate taxation of a significant digital presence, COM(2018) 147 final, Brussels, 21.3.2018.

[11] Communication from the Commission to the European Parliament and the Council, Business taxation for the 21st Century, COM(2021) 251 final, Brussels, 18.5.2021.

[12] *Ibid.*, p. 5.

[13] *Ibid.*, p. 8.

[14] Pillar Two would have an impact on for example the Anti-Tax Avoidance Directive, Council Directive (EU) 2016/1164 of 12 July 2016 laying down rules against tax avoidance that directly affect the functioning of the internal market. Amended by Council Directive (EU) 2017 of 29 May 2017), and the Interest and Royalties Directive, Council Directive 2003/49/EC of 3 June 2003 on a common system of taxation applicable to interest and royalty payments made between associated companies of different Member States.

[15] Press release by EU Commissioner Gentiloni dated 13 October 2021, and a news bulletin from Popa, Oana, European Union, IBFD, 14 October 2021 ("Commissioner Gentiloni welcomes G20's endorsement of agreed global tax reform and highligths EU plans to implement OECD pillars).

[16] Article 115 of the Treaty on the Functioning of the European Union.

Hungary would sign the agreement. In the final rounds of discussion, however, they all decided to enter into the agreement.[17] Accordingly, at present there seems to be no member state of the EU that is likely to veto any measures subsequently adopted by the EU.

As far as I can deduce from the list of countries which have entered the OECD/G20 Inclusive Framework agreement, it is only Kenya, Nigeria and Sri Lanka that have not entered into the agreement.[18] The likely reason is that they consider the agreement to be unfavourable to developing countries, not least when they have a large population constituting a considerable market for MNEs.

Following the decision by the OECD/G20 Inclusive Framework on Pillars One and Two on the 8 October 2021, a number of European states, including also the United Kingdom, have stated that they will allow a gradual termination of their unilateral taxes on digital services.[19] This is because the agreement on Pillar One and Pillar Two requires the removal of all digital services taxes.

The agreement and the underlying proposals are new, and the academic literature is still not that large. In this article I focus the discussion on the primary source, the 2020 OECD/G20 Pillar One Blueprint, and on what was included in the agreement on 8 October 2021.[20] I refer to the 8 October 2021 agreement as the "agreement", which is based on the 2020 OECD/G20 proposal contained in the report "Tax challenges arising from digitalisation – Report on Pillar One Blueprint", which I refer to as the "Pillar One Blueprint" or the "proposal". The October 2021 "agreement" is only a few pages with a general description of the rules and fundamental thresholds, whereas the "Pillar One Blueprint" covers

[17] *Financial Times*, 8 October 2021, OECD close to final deal on corporate tax.

[18] Cf. *Financial Times*, 1 July 2021, World's leading economies agree upon global minimum corporate tax rate.

[19] States abolishing digital services taxes include Austria, France Italy, Spain and the United Kingdom. In response the United States will terminate trade actions on France, which were put in force because of France's unilateral digital services tax. See De Lillo, Francesco, Report (IBFD), 25 October 2021, France joins agreement on transition from digital services tax to new international tax framework.

[20] OECD, Statement on a two-pillar solution to address the tax challenges arising from the digitalisation of the economy, Statement issued 8 October 2021 (available on www.oecd.org).

224 pages.[21] It should be emphasized that it is not clear what the details of the October 2021 agreement includes, and to what extent the details of the Pillar One Blueprint, will also be included in the forthcoming proposals for new legislation.

The work on the proposals for Pillar One and Pillar Two has been administered and largely conducted by the OECD. Staff at the OECD has worked together with experts from the governments of the member states of the OECD and of the large number of states and jurisdiction. These states are part of the "OECD/G20 Inclusive Framework on Base Erosion and Profit Shifting", which is often referred to as the "OECD/G20 Inclusive Framework" or just the "Inclusive Framework". Today there are 38 member states of the OECD, among them Sweden, which has been a member since the organization was established in 1961. Following the world financial crisis in 2008, the OECD and G20 started an ambitious work against the erosion of the corporate tax base, called work against "Base Erosion and Profit Shifting" or "BEPS". To the surprise of many, this project has been successful, and has led to much new legislation both at the national level and at tax treaty level. The Inclusive Framework continues the work of the BEPS project and includes almost 140 states and jurisdictions. A major task for the Inclusive Framework is to assess the implementation of the legislation developed from the BEPS project, and to work on the continuation of the project with Pillar One and Pillar Two.

At present, it is still unclear to what extent the 2020 Blueprint for Pillars One and Two ultimately will be followed. As regards Pillar One, it remains to be seen to what extent the distinction between "automated digital services" and "consumer facing business" will be upheld. Irrespective of the final outcome in that regard, some kind of digital tracing of consumers will be required in order to identify the residence state of the consumer.

[21] The leaders of the G20 met in Rome on 30 and 31 October 2021. At the meeting, the G20 endorsed the 8 October 2021 agreement by the states and jurisdictions forming the OECD/G20 Inclusive Framework. See Agianni, Vasiliki, Report (IBFD), 1 November 2021, G20 leaders welcome historic OECD/G20 Inclusive Framework global tax deal.

2 Traditional concepts for taxing businesses income

The basic structure for business taxation in an international environment was developed in the 1920s. After the First World War many states found it necessary to severely raise tax revenue. There was a need to rebuild society after the world war and raising tax revenue was considered necessary. An effect was that companies doing business in different jurisdictions faced the problem of paying considerable taxes, if the income was taxed both in the home state of the company, and in the state where its goods or services were sold. This is the problem of international (juridical) double taxation.[22]

The League of Nations was established after the First World War, and it had its seat in Geneva, Switzerland. The League of Nations identified the problem with international (juridical) double taxation and developed a model for an international instrument – the double taxation convention for the elimination of double taxation. This model contained several basic concepts for allocating income between states. It should be noted that it dealt not only with companies, but also with income earned by private individuals. The work of the League of Nations was later adopted by the OECD which has worked with international tax law issues since the organization was established in the early 1960s. The OECD has further developed the model tax treaty developed by the League of Nations, but the OECD model still contains the same basic framework as the first League of Nations model from 1928.[23]

In general, one can say that a tax treaty allocates the income earned by a company between the residence country and the other country, which

[22] In principle, double taxation has two forms: international juridical double taxation and economic double taxation. Tax treaties primarily deal with international juridical double taxation. It can be described as the situation when one taxpayer is taxed on behalf of the same income in at least two states for the same time period. Economic double taxation is the taxation of income, for example business profits, first at the corporate level and then in the form of dividend income (or capital gains) at the shareholder level. Other forms of income may also be subject to economic double taxation, such as interest and royalties.

[23] The 1928 Model tax treaty developed by the League of Nations was partly based on a report from 1923 by four academic tax experts, G.W.J. Bruins et al., League of Nations Econ. & Fin. Comm., Report on double taxation: Submitted to the Financial Committee, League of Nations Doc. E.FS.73.F.19 (1923).

usually is the source country of the income. This expresses the residence principle and the source country tax principle. For companies, residence is mostly identified either according to the registration principle or the seat principle. Sweden applies the residence principle, Chapter 6, Section 3 of the Income Tax Act (ITA). Companies are subject to an unlimited tax liability on its global income, if they are registered in Sweden. Non-resident companies are subject to tax in Sweden if they have income that is connected to a permanent establishment in Sweden. A withholding tax (30 per cent) applies on outbound dividend payments. In general, both the residence principle and the source country tax principle rely on some form of physical presence in order to apply. However, the registration principle concerning companies does in fact not rely that heavily on a physical presence, whereas the real seat theory does. It is not possible to deal with that in further detail in this article.

Another fundamental concept for traditional corporate taxation is the arm's length principle. It concerns how to determine the correct price on goods and services sold between related companies in different states. In brief, one can say the arm's length price should be the market price, namely the price that would have been used if the same or similar goods or services were sold between independent parties. Without the arm's length principle, it would be possible for multinational groups of companies to create extra costs in high-tax jurisdiction, and extra profit in low-tax jurisdictions.

The basic concepts developed by the League of Nations were implemented in tax treaty law, and thereby also had an influence on purely domestic law. The effect of the tax concepts, was more of a loose-legal binding effect, however, with considerable impact. The agreement now being reached within the OECD/G20 Inclusive Framework project has more of a hard-law approach, and it will take effect both in tax treaty law and in purely domestic law.

Following the financial crisis, which began in 2007–2008, many states became in urgent need of strengthened public finances, and an effective tax system to support this. From a European perspective, states like Greece, Italy and Spain had weak finances, but the problems were identified also on a broader scale. The OECD initiated a project regarding the preservation of the corporate tax base, called BEPS, which stands for: Base Erosion Profit Shifting. It concerns the threat of the corporate tax base being diverted to low-tax jurisdictions around the globe. The first reports in the BEPS project were published in 2012 and 2013.[24] They were followed up by additional reports on 15 different areas of tax law, which were submitted in 2015. One area was the taxation of the digital economy. After 2015, the OECD continued its work on new tax rules, not the least concerning the field of digital business. Numerous states unilaterally introduced measures taxing digital businesses. In focus for those laws, at least partly, were large US multinational companies. The acronym GAFA became well-known in the French tax debate and identifies tax legislation aimed at profits earned by Google, Apple, Facebook and Amazon. These new tax laws irritated the US government under President Trump, and lead to some counter measures issued by the US on goods sold by French companies in the US.

In general, the OECD reports on the taxation of digital business identified a risk that the profits earned by the giant tech companies turned out to be un-taxed or at least taxed at very low rates according to the traditional tax principles.[25] When targeting a specific market, these companies did it through companies established in low-tax jurisdictions such as the Republic of Ireland. The products are sold digitally in other states, but due to use of digital services there is no need for any physical presence in the market state. According to traditional tax principles, there is accordingly no possibility for the market state (source state) to tax the income generated in that state.

As already mentioned, several states have unilaterally introduced tax legislation for taxing the profits of such digital businesses. Previously, the European Commission has also issued proposals for the taxation of digital business. However, in the recent year a fundamental proposal for the taxation of digital businesses has been put forward by the OECD and endorsed by the G20, most of the EU member states, and also a large number of other states such as the Peoples' Republic of China, India, Russia, Brazil and South Africa. Moreover, the United States under the new Biden administration and with the finance minister Jane Yellen, strongly endorsed the proposal. Sweden is also behind the proposal, even if it seems to be with some hesitation: Sweden is dependent on digital

[24] OECD, Addressing base erosion and profit shifting, Paris, 2013, and OECD, Action plan on base erosion and profit shifting, Paris, 2013.

[25] On the problems with digitalization of business and the erosion of corporate taxation, see OECD, Addressing the tax challenges of the digital economy, Action 1: 2014 Deliverable, 2014, Chapters 5 and 6.

businesses, and there is a risk that some of the corporate profits will be taxed in the market states, which for many Swedish companies will be in other states than in Sweden. With a population of only 10 million inhabitants, Sweden is a small market, and the allocation of corporate profits to Sweden according to the new legislation will probably be small. The European Commission has already made it clear that it is behind the OECD/G20 Inclusive Framework proposals and agreement, and that it intends to propose to the EU to adopt them, when necessary, through directives or changes to existing directives.

3 The OECD/G20 Inclusive Framework agreement and proposal according to Pillar One and Pillar Two

The OECD/G20 Inclusive Framework proposal and agreement on Pillar One and Pillar Two includes fundamental changes to the system for corporate taxation. Pillar One, which is discussed in this article, concerns a new model for taxing corporate income stemming from digital business. Pillar Two deals with additional rules to supplement the traditional principles for taxing corporate income. The idea is to make the traditional principles more protected in a global economy, and that they should be used in addition to the new principles concerning digital businesses. An important aspect of Pillar Two is that, in effect, it enforces a minimum corporate tax rate of 15 per cent.

The Pillar One proposal issued by the OECD a year ago, in October 2020, is complex, full of details, and there are still many issues that needs to be resolved by the participating states.[26] It is not possible to go into all the details of the proposal in this article. Instead, I will focus on some aspect of the definitions of the digital business that are suggested will be covered by the new tax regime.

What are the driving forces behind the proposal by the OECD and the Inclusive Framework ("IF") states? One important force is the risk that major tech companies, active on a global basis, will earn income that will largely remain untaxed. The traditional set of corporate tax rules, with

[26] In academic literature proposals have been made to simplify Pillar One, see for example Graetz, Michael J., A major simplification of the OECD's Pillar 1 proposal, Tax Notes Federal, January 11, 2021, pp. 213–225.

their emphasis on physical presence, is easy to circumvent, or, rather, they do not adapt to the form of business that the tech companies undertake. Another driving force is the tax interest of the market state, that is the state where the consumer is resident. Many of the business models conducted by tech companies include the involvement of consumers: User postings on Facebook and Youtube are only two obvious examples.[27]

The proposed corporate tax rules for large tech companies will work in parallel with the traditional corporate tax systems, which probably is one of the major hurdles with the proposal.

As previously mentioned, on October 8, 2021 an agreement was made by most of the states and jurisdictions participating in the OECD/G20 Inclusive Framework on both Pillar One and Pillar Two. Regarding Pillar One, which is discussed in this article, agreement was made on the following – important – details.[28] The in-scope companies are multinational enterprises (MNEs) that have a global turnover of more than 20 billion euros. They should also have a profitability that exceeds 10 per cent. The profitability is calculated as profit before tax divided with revenue. After 7 years, the turnover threshold is reduced from 20 billion euros to 10 billion euros. The October 2021 agreement explicitly excludes "extractives" (which presumably includes mining activities as well as oil and gas extraction) and regulated financial services from the scope of the proposed tax system.

A further condition (called "nexus") requires that an in-scope MNE has at least a profit of 1 million euros from a market jurisdiction in order for that jurisdiction to be allocated a taxing right according to Amount A, which is the income from digital business discussed in this article. For a

[27] A general description of the digital economy is outlined by the OECD in the report, Addressing the tax challenges of the digital economy, Action 1: 2015, Final report, Paris, 2015 (esp. pp. 51–74). A critical study of the digital economy and the tech-giants is made by Harvard professor (of business administration with an academic background in social psychology) Shoshana Zuboff in her bestselling book, The age of surveillance capitalism. The fight for a human future at the new frontier of power, Profile Books, London, 2019. It is an interesting book, even if I think that she to some degree overemphasises problems with the digital economy and overlooks the possibilities and benefits it has brought to both producers and consumers.

[28] OECD, Statement on a two-pillar solution to address the tax challenges arising from the digitalisation of the economy, Statement issued 8 October 2021 (available on www.oecd.org).

jurisdiction ("market state") with a GDP lower than 40 billion euros, the nexus will be met at a revenue of 250,000 euros from that jurisdiction.[29]

The income that will be sourced to the market state and taxed there, is called "quantum". It will constitute 25 per cent of the residual profit which exceeds 10 per cent of the revenue earned by the MNE at issue. A revenue allocation key will be used to divide the taxable income between different market states. The basic approach is that the income will be allocated to that market state where the final consumers are resident.

Sourcing rules are important to calculate Amount A.[30] It is through sourcing rules that the consumers are identified and thereby the "market state", which will be allocated a taxing rights according to Pillar One. It is not possible to discuss the details of the proposal in this article. However, a few remarks should be made. When it comes to digital services, which primarily are covered by the term "automated digital services" complex mechanisms are required for identifying the consumer. For example, when it comes to online advertising services, it is the real-time location of the viewer that constitutes the "sourcing rule", which will identify the market state.[31] In order to practically apply this sourcing rule, a number of different indicators are used. They include the geolocation of the device, the jurisdiction of the IP address, and other available information.[32] Another example is the sale or alienation of user data. The sourcing rule is the jurisdiction of the real-time location of the user that is the subject of the data being transmitted, at the time when the data was collected.[33] Among the relevant indicators are the jurisdiction of the geolocation of the device of the user at the time of collection, and likewise, the jurisdic-

[29] The five countries with the highest GDP in 2020 were 1) The United States (20,936 billion USD), 2) China (14,722 billion USD), 3) Japan (5,064 billion USD), 4) Germany (3,806 billion USD), and 5) The United Kingdom (2,707 billion USD). Sweden ranked 22 with a GDP of 537 billion USD in 2020. Examples of countries close to the threshold of a GDP of 40 billion euros were Cameroon, Tunisia, Bahrain, Uganda, Bolivia and Paraguay. A large number of countries have a annual GDP well below 40 billion euros. The statistics are obtained from the World Bank (www.worldbank.org). The World Bank statistics use USD as currency, and the Pillar One thresholds are set in euros. At the time of writing (2 November 2021) 1 euro equals 1.16 USD (currency information from Svenska Handelsbanken, www.handelsbanken.se).

[30] OECD, Pillar One Blueprint, 2020, Chapter 4.

[31] OECD, Pillar One Blueprint, 2020, para. 238.

[32] OECD, Pillar One Blueprint, 2020, para. 239.

[33] OECD, Pillar One Blueprint, 2020, para. 243.

tion of the IP address.[34] User profile information and billing address are also used as indicators. It is the MNE that has to provide the information needed to apply the sourcing rules. However, it should not be done at the level of the individual consumer. It is suggested that this can be obtained from "systemic data" kept by the MNE, under the assumption that it has a "robust internal control framework on which the tax authorities can rely".[35] It remains to be seen whether this can be achieved.

The taxable income will be calculated using financial accounting.[36] From a Swedish perspective, this is another breach with traditional concepts, because special tax rules normally apply, although they may be influenced by financial accounting.

Double taxation will probably be a major problem with the new rules. The reason is that they will apply in parallel with the traditional tax rules.[37]

It is highly likely that the new tax rules will be difficult to interpret and apply. The EU is planning to issue two directives, one for each of the two pillars agreed upon by the OECD/G20 and the Members of the Inclusive Framework.[38] According to the proposal for Pillar One, there will be new mechanisms for dispute prevention and resolution mechanisms.[39] As a means of last resort, the regular administrative system of each participating state may be used to interpret the new rules and resolve tax disputes.

How is the shift of tax revenue to the market state justified? The tax debate on this issue identifies several reasons. One reason is that the digital economy with little need for physical presence in the market state has made the traditional tax principles obsolete. They were to a high degree focused on corporate residence and physical presence through, for example, offices, factories, and personnel. In the digital economy services and goods are provided online, and it is possible to reach a high degree of integration in the economy of the market state without any physical presence. Another reason is that consumers to a high-degree integrate with

[34] OECD, Pillar One Blueprint, 2020, para. 244.

[35] OECD, Pillar One Blueprint, 2020, para. 388–391 (quote from para. 390).

[36] OECD, Pillar One Blueprint, 2020, para. 407–410.

[37] OECD, Pillar One Blueprint, 2020, chapter 7.

[38] European Commission, Communication from the Commission to the European Parliament and the Council. Business taxation for the 21st century, COM(2021) 251 final, Brussels, 18.5.2021.

[39] OECD, Pillar One Blueprint, 2020, chapter 9. The issue of dispute prevention and resolution are addressed under the title "Tax Certainty".

other consumers and the supplier of digital services, and thereby participate in creating corporate value.[40] Therefore it is justified for the market state to tax the earning from those services. The market state has for example provided the infrastructure for the digital market penetration. That infrastructure can include obvious parts like the availability of telecommunication facilities and fiber optic cables. From a larger perspective one could add educational level and computer knowledge supplied by the educational system of the market state.[41]

4 Digital business activities to be covered by the proposed tax legislation

4.1 Introduction

There are two major digital business activities that are suggested that they should be covered by the new tax. This income will be taxed according to what is labelled as “Amount A”. There is also an “Amount B” that will cover “baseline marketing and distribution activities”, that is a category of income that many tech companies have, and which can give rise to considerable problems with current rules. It should however be noted that regarding “Amount B” there is only a revision of details of the current rules, the principal approach – including the arm’s length principle – is maintained.

Regarding “Amount A” there is, a fundamentally different approach for allocating the taxable income between states. This approach is not new in itself, though it has not been used to any wider extent on an international level. The tax theoretical label is the formulary apportion

[40] There has been much academic criticism on “value creation” as a theoretical ground for allocating taxing jurisdiction to the market state. In brief, the concept is considered too vague. See for example, Hey, Johanna, “Taxation where value is created” and the OECD/G20 Base Erosion and Profit Shifting initiative, Bulletin (IBFD), 2018, pp. 203–208, and Schön, Wolfgang, Is there finally an international tax system?, World Tax Journal (IBFD), 2021, pp. 357–384.

[41] Cf. on this topic Li, Xiaorong, A potential legal rationale for taxing rights of market jurisdictions, World Tax Journal (IBFD), 2021, pp. 25–61.

method, and it has for decades been rejected by the OECD.[42] However, times are changing.

There are two general fields of activity that are covered by the "Amount A". They are "Automated digital services" ("ADS") and "Consumer facing business" ("CFB"). The proposed tax rules will tax net profits generated by such digital business activities. The income identified as "Amount A", will be allocated between different states on behalf of the nexus to the different states. A considerable part of that income will be allocated as taxable income to the state in which the digital business has its market.

4.2 "Automated digital services" or "ADS"

In general, "automated digital services" refers to MNE activities that have provided digital services all over the world with little or no infrastructure in their market states. In addition, their business models include interaction with customers, who also provide content to the digital services and increase the economic value of the services provided.[43] The fact that consumers provide value to the services, for example postings on Facebook or videos on Youtube, is a key rationale for arguing that the market state should have a taxing right on the income generated by the MNE's digital business.

What constitutes an ADS is identified in three steps. First, according to a positive list of activities that are considered to be ADS. Second, with a negative list of activities that are not considered to be ADS. Third, with a general definition which apply on activities that are not covered by either the positive or negative list. The practical application is to begin with the positive list, and then continue with the negative list, and ultimately, if no answer is provided in the first two steps, apply the general definition.

In order to get a general understanding of ADS, it is suitable to begin with the general definition. It stipulates that an ADS is:

[42] OECD, Transfer Pricing Guidelines, Paris, 2017, para. 1.16–1.32. In these guidelines, the OECD explicitly rejects global formulary apportionment, which actually is what the taxation of Amount A is. Much has happened since the recent version of the guidelines was published in 2017.

[43] OECD, Pillar One Blueprint, 2020, para. 24.

- automated, which means that when the system is established the provision of the service to a particular user requires minimal human involvement on the part of the service provider, and
- digital, which means that it is provided over the Internet or an electronic network.[44]

The meaning of "automated" is at the centre of the definition. It implies that it is possible for the consumer to use the service through different kinds of equipment, such as computers and digital communication, without interaction with personnel employed by the provider. It is also possible for the providing company to scale up its business to meet a higher demand, with "minimal human involvement".[45]

The "positive list" on ADS gives a good picture on the activities that are covered. The positive list includes:

- Online advertising services,
- Sale or alienation of user data,
- Online search engines,
- Social media platforms,
- Online intermediation platforms,
- Digital content services,
- Online gaming,
- Standardised online teaching services, and
- Cloud computing services.[46]

It is recognized that these categories are not mutually exclusive, and that there may be an overlap between them.[47]

The negative list contains five categories of activities, namely:

- Customised professional services,
- Customised online teaching services,
- Online sale of goods and services other than ADS,
- Revenue from the sale of goods of a physical nature, irrespective of "network connectivity", which includes "the Internet of things", and

[44] OECD, Pillar One Blueprint, 2020, Box 2.1, p. 23.

[45] OECD, Pillar One Blueprint, 2020, Box 2.1, p. 24.

[46] OECD, Pillar One Blueprint, 2020, para. 44.

[47] OECD, Pillar One Blueprint, 2020, para. 45.

– Services providing access to the Internet or other form of electronic network.[48]

There is a major difference between the activities on the positive list and those on the negative list. The positive list contains activities that are "automated", and the negative list contains activities that are designed for a particular customer and that includes some form of direct human involvement.

There is an interesting example from the perspective of the legal profession on "customised professional services". It is recognized in the report that law firms rely heavily on AI when it comes to due diligence activities. A due diligence may for example be performed when one company plans to buy another company ("target company"). In order to investigate the activities of the target company a law firm is hired to make an inquiry, which for example includes contracts and tax matters. The law firm may use AI in order to identify issues of interest. This is considered to be a "customised professional service" that falls out of scope of ADS. The reason is that human involvement is necessary to develop the AI and to evaluate the results it provides. However, the payments that the law firm makes to the provider of the AI developer may be covered by ADS, according to the OECD report. The reason is that the activities of the AI developer can constitute cloud computing or digital content service, see the previous discussion on the "positive list".[49]

It is of course possible that the large MNEs that are at issue, can have, or even frequently, will have parts that deal with ADS, and parts that will not. This is a classic issue in tax law: Should one tax according to the different parts or consider the activities of the corporate group as a whole?

If the ADS parts are clearly identifiable the "revenue streams" from those activities should be taxed separately. However, if the ADS parts are highly integrated in the parts of other non-ADS business activities, they should be considered as a whole. Only if the ADS part forms the substantial part, should the business activity as a whole be considered as ADS. If the ADS is a smaller part of the integrated business activity, it should not as a whole be considered as ADS. It goes without saying that the distinctions will be difficult to generally lay out in tax law, and legal practice concerning specific business activities will be necessary. However, this is

[48] OECD, Pillar One Blueprint, 2020, para. 46.

[49] OECD, Pillar One Blueprint, 2020, Box 2.22, para. 32–33.

something with which tax law is familiar, and it will be possible to do it in case law. Of course, for the benefit of tax payers and tax agencies, attempts to draw the borders in tax law provisions are welcome.

4.3 Consumer facing business ("CFB")

The other major category of income covered by Pillar One is "consumer facing business" ("CFB"). It includes a large area of business activities. The general description is that CFB is business that generate income from the sale of goods and services, which commonly are sold to consumers.[50] Accordingly, goods and services commonly provided to other businesses are not covered. There is of course a wide range of such activities, and I would assume that equipment and maintenance of, for example, industrial facilities, ports, airports and windmills would not be covered. As always in tax law, there will likely be a number of border line situations which have to dealt with in case law.

Even if the scope of CFB is large, some activities within the scope are explicitly discussed in the OECD report. One example is pharmaceuticals, which are covered if they are sold to consumers. Not least from the current pandemic, a number of global pharmaceutical companies have provided the world with vaccines. The pandemic is not addressed in the OECD report, but from my understanding it seems a bit unclear whether vaccine producing companies, or that part of the multinational pharmaceutical enterprises, would be covered. From my understanding, it has not been the case that consumers themselves have purchased the vaccines, but governments have done so and distributed them through the health care system in their respective countries. The national health care system may contain private health care providers, but that would not make any difference. The vaccines are not commonly sold to consumers, which is the general prerequisite for pharmaceuticals to be covered by CFB.[51] Therefore, it seems unlikely, at least according to the current version of the proposal for defining CFB, that vaccine producing companies would be covered on behalf of that activity.

It is recognised that some goods and services may be of dual use, that is, of use for both consumers and other businesses. Cars, computers and some medical products (such as blood pressure monitors) are examples

[50] OECD, Pillar One Blueprint, 2020, para. 52.

[51] OECD, Pillar One Blueprint, 2020, para. 57.

of products fall within this category, according to the OECD.[52] The approach taken is that if the goods or service is commonly sold to consumers, then all sales of that goods or service will be covered by the definition of CFB. This is, evidently, an either-or-approach. Only if there is a marginal sale to consumers, will the sale be excluded from the scope of CFB.[53]

There are several exclusions and "carve-outs" from the scope of CFB. In my view, some of them follow from the general definition of CFB, but are still explicitly discussed in the report. The sale of natural resources is excluded. The meaning of "natural resources" is, of course, wide, and the OECD discusses for example agriculture, forestry, and the mining industries. To me it seems unlikely that the raw materials extracted from mining would meet the general definition of CFB, because it needs to be processed before reaching the consumer in some form. However, products from agriculture and fishing may not require that, and can more easily be provided directly to the consumer. The extraction and production of fossile energy and the production of renewable energy is also discussed, but the conclusions are vague. The issues are sensitive, and it seems that they will have to be further discussed.[54]

5 Concluding remarks

The agreement on Pillar One and Pillar Two by most of the more than 130 states participating in the OECD/G20 and Inclusive Framework project means a fundamental shift in the global system for taxation of corporate profits. The different thresholds for the new rules to apply, makes this specific tax system applicable only for very large MNEs. The European Commission plans to propose a directive, or amendments in current directives, for the implementation of at least Pillar Two.[55] Basically, there are two fundamental shifts if the two pillars ultimately come

[52] OECD, Pillar One Blueprint, 2020, para. 93.

[53] OECD, Pillar One Blueprint, 2020, para. 93. Practical aspects of drawing a line between sales to consumers or businesses seems to be the general reason for the approach (ib., para. 95–97).

[54] OECD, Pillar One Blueprint, 2020, para. 116–121.

[55] In this article I have focused on Pillar One. Regarding Pillar Two, there are problems with the compatibility with fundamental freedoms of EU law cf. Brokelind, Cécile, An overview of legal issues arising from the implementation in the European Union of the OECD's Pillar One and Pillar Two Blueprint, Bulletin (IBFD), 2021, pp. 212–219.

into effect. First, there will now be a global set of rules for the taxation of corporate profits. It is possible that these rules will apply in more than 130 states and jurisdictions all over the world. Second, the new rules include a new approach for dividing the tax base between companies, namely a formulary apportionment method that allocates parts of the income to the market state.

Even if it is only large MNEs that will be covered, I consider it likely that the approach will permeate into the system for taxing other (including small and medium-sized) companies conducting international business activities. The directives on corporate taxation which the EU has introduced on certain cross-border situations, have had such spill-over effects on strictly internal situations in the Member states. For example, the implementation of the EU Merger directive had such effects in Sweden on internal re-organizations of corporations and corporate groups.

Will Pillar One and Pillar Two ultimately take effect? So far, it is only the governments of most of the participating states and jurisdictions within the project that have reached an agreement. For the agreement to take effect, it will have to be implemented with national law and tax treaty law. It remains to be seen how national parliaments will react on the agreement. That is a problem. The proposals on Pillar One and Pillar Two have, so far, been a project discussed by governments and the large MNEs through different lobby organizations. There is also an academic debate, however, in a rather limited group. In my view, there has so far not been any public debate.

Individual rights and the protection of personal integrity is likely to be an issue when the details of Pillar One will be put forward as proposals for new national tax legislation. The earlier discussion of Amount A makes it clear that the location and consumption patterns of billions of individuals will form the basis for allocating taxing power to the market state. The issues of individual rights and personal integrity are hardly discussed in the OECD Pillar One Blueprint. It still remains to be seen what the new legislation, for example, an EU Directive on Pillar One will contain in this regard. When individual rights and personal integrity is discussed in relation to the digitalisation of society, references are often made to the oppressive state portrayed by George Orwell in his novel "1984". There is, however, a precursor to this novel, which also portrays a state that keeps its citizens under close surveillance, namely Evgeny Zamiatin's novel "We" (*Мы*). In the novel, we follow an engineer (with the impersonal name "D-503") who on the commission of the state and its leader, the Great Benefactor, is developing a new technological tool that will keep the citizens under an even stricter control. The technological tool is called "Integral". In a famous passage, the engineer formulates how this totalitarian state considers the rights of the individual, a way of reasoning which also provides the title of the novel:

> "There are ideas made of clay, and there are ideas sculpted for the ages out of gold or out of our precious glass. And to determine what material an idea is made of, all you have to do is let a drop of powerful acid fall on it. Even the ancients knew one such acid: *reductio ad finem*. That's what they seem to have called it. But they were afraid of this poison. They preferred to see at least some kind of heaven – however clay, however toylike – to this blue nothing. But we are grown-ups, thanks be to the Benefactor, and don't need toys.
>
> Look here – suppose you let a drop fall on the idea of 'rights'. Even among the ancients the more grown-up knew that the source of right is power, that right is a function of power. So, take some scales and put on one side a gram, and on the other side a ton; on one side 'I' and on the other 'We', OneState. It's clear, isn't it? – to assert that 'I' has certain 'rights' with respect to the State is exactly the same as asserting that a gram weighs the same as a ton. That explains the way things are divided up: To the ton goes the rights, to the gram the duties. And the natural path from nullity to greatness is this: Forget that you're a gram and feel yourself a millionth part of a ton."[56]

It is important to follow the forthcoming legislation on Pillar One (and, of course, Pillar Two) and identify eventual infringements on individual rights. It is a pity that this issue has hardly been addressed in the materials so far published. Considering that there has not been any public debate, and that multilateral treaties and domestic legislation are intended to be decided in 2022, to take effect in 2023, there is not much time. There is a potential problem with consumer integrity in the huge amounts of information that will be required in order to identify the market state.

[56] Zamyatin, Yevgeny, We, Penguin, London, 1993 (1924), p. 128 (translation to English by Clarence Brown). On the fascinating life of Zamyatin, see the biography of the Oxford professor of Russian literature, Julie Curtis, The Englishman from Lebedian' – A life of Evgeny Zamiatin (1884–1937), Ars Rossica (ed. David Bethea), Academic Studies Press, Brighthon (MA, USA), 2013. There is also a Swedish translation of the novel, Zamjatin, Jevgenij, Vi, Modernista 2015 (1959), translation by Sven Vallmark, including a foreword by Nils Håkanson.

Liane Colonna

The AI Regulation and Higher Education: Preliminary Observations and Critical Perspectives

1 Introduction[1]

The introduction of Artificial Intelligence (AI) into educational contexts may be traced to the 1970s, when researchers were interested in understanding how computers could substitute one-to-one human tutoring.[2] While the development of AI-powered teaching and learning tools has steadily progressed, Higher Education (HE) institutions have been slow to adopt them. However, the Covid-19 pandemic, has drastically changed the landscape, forcing universities to rely on technology for virtual learning. Long gone are the days of clunky desktop computers sitting in lonely student computer labs. These are the days of virtual and

[1] The support of The Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS), Ethical and Legal Challenges in Relationship to AI-driven Practices in Higher Education (MMW2020.0138), is gratefully acknowledged. Additionally, I would like to express my gratitude to Professor Teresa Cerratto-Pargman, Associate Professor Cormac McGrath and Professor Cecilia Magnusson Sjöberg who have provided helpful comments on the manuscript, or parts thereof, at various stages in its development. An additional thanks goes to the editors of this volume, Assistant Professor Katja de Vries and Professor Mattias Dahlberg for their very constructive feedback. As far as shortcomings are concerned, they are all attributable to the author.

[2] B.S. Bloom, *The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring*, 13 Educational Researcher 4 (1984).

augmented realities, remote-based proctoring, Ed Tech robots, predictive learning analytics, and more.[3]

While AI may present incredible opportunities to improve teaching and learning and could have a huge impact on the future of education[4], it poses new and far-reaching ethical, legal, and social challenges.[5] Until recently, the rapid development of technology in this context has generally outpaced policy debates and regulatory frameworks about how best to develop and use AI in HE in ways that are not just equitable, ethical, and effective but also just, fair, and caring.[6] That said, on 21 April 2021, the European Commission published a legislative proposal for a "Regulation on a European Approach for Artificial intelligence" (the AI Regulation) which expressly addresses the use of AI in the educational context.

This paper investigates the proposal from the perspective of HE. More specifically, it seeks to make some preliminary observations as well as offer some critical perspectives concerning the way the proposed AI Regulation addresses the educational context, particularly in HE. It explores whether the proposal is old wine in new bottles or representative of a fundamental shift in approach towards the more ethical use of AI in the HE.

2 Background to the AI Regulation

The proposal was the result of a broad consultation process and many years of investigating whether AI requires specific regulation, and if so, how normative assumptions and ethical principles should be reflected

[3] R. Huang, J.M. Spector, Junfeng Yan, *Introduction to Educational Technology*, In: Educational Technology: Lecture Notes in Educational Technology (eds. Ronghuai Huang, Kinshuk, Mohmed Jemni, Nian-Shing Chen, J. Michael Spector) (Singapore, Springer 2019).

[4] *But see* Cerratto Pargman T. and Cormac McGrath, *Be Careful What You Wish For! Learning Analytics and the Emergence of Data-Driven Practices in Higher Education*, In: Digital Human Sciences (ed. S. Petersson)(Stockholm, Stockholm University Press 2021) (discussing the "techno-romanticism" and hype surrounding learning analytics), https://www.stockholmuniversitypress.se/site/chapters/e/10.16993/bbk.i/#disqus_thread.

[5] Cerratto Pargman T. and Cormac McGrath, *Be Careful What You Wish For! Learning Analytics and the Emergence of Data-Driven Practices in Higher Education*, In: Digital Human Sciences (ed. S. Petersson)(Stockholm, Stockholm University Press 2021), https://www.stockholmuniversitypress.se/site/chapters/e/10.16993/bbk.i/#disqus_thread.

[6] P. Prinsloo and S. Slade, *Big data, Higher Education and Learning Analytics: Beyond Justice, Towards an Ethics of Care*, In: Big Data and Learning Analytics in Higher Education, (Springer 2017), 109–124.

in the law. It builds on key documents such as the AI HLEG, Ethics Guidelines[7] which sets forth ethical imperatives in the context of AI and the Commission's White Paper on Artificial Intelligence[8] which proposes a risk-based regulation for AI with sector and application-specific risk assessments and requirements as opposed to blanket requirements or bans. The aim of the initiative is to establish a comprehensive, "futureproof" legal framework regulating AI in all sectors, including education, to foster economic growth, to ensure that there is harmonization of approaches between all 27 of the EU Member States, to offer safety and legal certainty to both consumers and industry and to create responsible[9] and trustworthy[10] AI.

As a legislative proposal, a debate and an approval process will follow, which might last until 2022. It will likely be amended by members of the EU Parliament as well as by governments of each EU Member State. Even after its publication in the Official Journal of the European Union, its implementation is likely to be incremental with full application after 24 months.[11]

3 The regulatory challenge

In a recent report by the European Parliament on AI in education, culture and the audiovisual sector, released after the AI Regulation in May 2021, it was noted: "Whilst it is easy to understand the potential effects of AI on sectors such telecommunications, transportation, traffic management, health care, evaluating its long-term effects on education" is "considera-

[7] Independent High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI 15 (Apr. 8, 2019), at https://ec-europa-eu.ezp.sub.su.se/newsroom/dae/document.cfm?doc_id=60419.

[8] *White Paper on Artificial Intelligence - A European Approach to Excellence and Trust* 16, European Commission (February 19, 2020), http://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[9] For more on responsible AI, see Virginia Dignum, *Responsibility and Artificial Intelligence*, In: (Eds. Markus D. Dubber, Frank Pasquale, and Sunit Das) The Oxford Handbook of Ethics of AI (Oxford, 2020).

[10] For more on trustworthy AI see Luciano Floridi, *Establishing the Rules for Building Trustworthy AI*, 1 Nature – Machine Intelligence 261 (2019).

[11] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final) (hereafter 'AI Regulation'), Article 85(2).

bly more challenging."[12] It further noted: "The potential impact of AI on education, culture and the audiovisual sector" is "rarely discussed and is mostly unknown. Yet this question is of utmost importance because AI is already being used to teach curricula…"[13] Here, it is interesting to note that the EU has decided to introduce the explicit regulation of AI in the educational sector without a full understanding of how AI impacts the sector. This, of course, highlights the regulatory challenge at hand.

There is a lack of confidence when it concerns the application of AI with cries from across almost all segments of society regarding the potential for the use of AI to lead to erroneous, opaque, brittle, and biased decisions. There is a fear that the use of AI can put safety, health and fundamental rights to privacy, data protection, free expression and assembly, non-discrimination, dignity at risk. For example, in the educational sector, students are concerned that marginalized students such as those with special needs and those from low-income families may disproportionately and unfairly have to pay the price of AI-based teaching and learning technologies, based on potentially racist, sexist, ableist, and hetero-centrist norms being reflected in the systems.[14] They are also deeply concerned about the role of automated individual decision making in HE, including profiling, where there is no human involvement such as where AI flags a student of color for cheating when the behavior is not actionable.[15] This could, for example, occur when an AI-based proctoring system identifies that a student is "cheating" but in reality all that has happened is that the student's child has entered the test-taker's environment to ask for a snack, causing the student to look away from the screen towards a second person.[16]

[12] European Parliament, Report on artificial intelligence in education, culture and the audiovisual sector (2020/2017(INI)), https://www.europarl.europa.eu/doceo/document/A-9-2021-0127_EN.html.

[13] European Parliament, Report on artificial intelligence in education, culture and the audiovisual sector (2020/2017(INI)), https://www.europarl.europa.eu/doceo/document/A-9-2021-0127_EN.html.

[14] Forthcoming, Liane Colonna, *Legal Implications of Using AI as an Exam Invigilator*, In: 2020–2021 Nordic Yearbook – Law in the Era of Artificial Intelligence (eds. Liane Colonna and Stanley Greenstein)(Stockholm, The Swedish Law and Informatics Research Institute (IRI)).

[15] Id.

[16] Id.

On one hand, there is a need to intervene before the application of AI within society acquires even greater momentum and AI-based products and services grow larger, more complex, and, naturally, more resistant to regulatory prodding.[17] However, if intervention comes to early than the EU may regulate without fully understanding the technology's likely impact. This dilemma is sometimes referred to as the Collingridge Dilemma[18], or more colloquially as the problem of "chasing a moving target." To put it differently, when it comes to the regulation of AI in the education sector, regulators face an "uncertainty paradox"[19], "where they must make decisions in the absence of reliable risk information or foreknowledge of technological developments."[20]

Basically, the EU's goal is to get ahead of the broad, unregulated use of AI and ensure that society knows that high-risk AI has gone through an extensive vetting process so that individuals can trust it. On the other hand, there is a visceral concern that if the EU starts to introduce cumbersome legislation, then it will fall behind countries like China and the United States. As Eric Schmidt, the former CEO of Google, has said: "The EU should be an "innovation partner to the US," in order to be able to compete with China."[21] Instead, "the EU did regulation first and I think that's a mistake."[22] Sometimes this debate is framed as "regulation

[17] Lyria Bennett Moses, *How to Think about Law, Regulation and Technology: Problems with "Technology" as a Regulatory Target*, 5 Law, Innovation and Technology 1 (2013).

[18] Anna Butenko and Pierre Larouche, *Regulation for Innovativeness of Regulation of Innovation?*, 7 Law, Innovation and Technology 52 (2015), 70. (According to the Collingridge dilemma, "if regulators want to achieve results, they should act early, but then the full range of risks and benefits is unknown, and if they wait until the risks and benefits are clear, the situation solidifies in a manner that makes it difficult and expensive to introduce regulatory changes." However, in the "early stages of technological development, there is insufficient information regarding potential harms and benefits.").

[19] Marjolein van Asselt, Ellen Voss and Tessa Fox, *Regulating Technologies and the Uncertainty Paradox*, In: Dimensions of Technology Regulation (eds. M. Goodwin, B. J. Koops, & R. Leener)(Wolf Legal Publishers 2010), 259–284.

[20] Lyria Bennett Moses, *How to Think about Law, Regulation and Technology: Problems with "Technology" as a Regulatory Target*, 5 Law, Innovation and Technology 1 (2013).

[21] Pieter Haeck, *Ex-Google Boss Slams Transparency Rules in Europe's AI Bill, Politico* (31 May 2021) https://www.politico.eu/article/ex-google-boss-eu-risks-setback-by-demanding-transparent-ai/.

[22] Id.

versus innovation" or "upstream governance versus permissionless innovation."[23]

The AI Regulation addresses the pacing problem as well as the need to balance the relationship between innovation and regulation with its risk-based approach. This approach entails that the regulation differentiates between uses of AI that create (i) an unacceptable risk (Title II), (ii) a high risk (Title III) (iii) a limited risk (Title IV) and (iv) a low or minimal risk (Title IX). The first category is generally prohibited, the second category is subject to compulsory regulation such as ex-ante conformity assessment, the third category is permitted but subject to transparency obligations, and the last category is only regulated by voluntary codes of conduct.[24] Thus, the EU takes the position that it is possible to update the law of the analog age (as well as the legacy of today's ICT applications that are already up and running within the digital information society, including HE!), without hindering innovation by focusing on high-risk AI and maintaining that the vast number of use cases are not subject to the regulation. It can be argued that this approach provides legal certainty: once the proposal is finalized, businesses will know the rules of the games so they can invest properly. The rules are based on globally acceptable principles, e.g., data quality, transparency, human oversight etc. so that any great shock to the business community should be avoided. The main difference is that it is no longer enough to just sign up for a list of principles – now AI providers must prove that they abide by them, at least for those systems that present risks to health, safety and fundamental rights.

In addition to the risk approach, the proposed AI Regulation also relies on techniques to monitor and update the legislation which may prove helpful to address the pacing problem. By avoiding the creation of a law that becomes stringently fixed and difficult to change, the AI Regulation allows for incremental adjustments in governance as needs arise.[25] One example is the provisions in Article 7, for the addition of new applications to the list of high-risk uses. This approach provides flexibility and

[23] *See e.g.*, Andrew McCafee, *EU Proposals to Regulate AI Are Only Going to Hinder Innovation*, Financial Times (25 July 2021), https://www.ft.com/content/a5970b6c-e731-45a7-b75b-721e90e32e1c.

[24] For more on the risk-based approach see Stefan Larson and Jonas Ledendal, "AI i offentlig sektor: Från etiska riktlinjer till lagstiftning" in this volume.

[25] Gregory N Mandel, *Regulating Emerging Technologies*, 1 Law, Innovation and Technology 75, 89 (2009).

adaptability as well as mentally prepares industry that the list of high-risk AI is not set in stone.[26]

4 The AI Regulation and Education[27]

Referring explicitly to the educational sector, Annex III of the AI Regulation states that AI systems used for "assessing students in educational training" constitutes high-risk AI.[28] It also refers to "AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions."[29] Recital 35 underlines that "AI systems used in education or vocational training[30], notably for determining access or assigning persons to educational and vocational training institutions or to evaluate persons on tests as part of or as a precondition for their education should be considered high-risk, since they may determine the educational and professional course of a person's life and therefore affect their ability to secure their livelihood."[31]

First, it is unclear whether the use of AI to predict (possibly falsely) the potential success of a student and then suggest to the student not to pursue a particular line of training would fall under the category of AI systems intended to be used for the purpose of determining access to educational institutions.[32] In other words, what if a student decides, based on a prediction made by AI, to drop out of an educational program,

[26] Id.

[27] Part of this section is based on previous work: Forthcoming, Liane Colonna, *Legal Implications of Using AI as an Exam Invigilator*, In 2020–2021 Nordic Yearbook - Law in the Era of Artificial Intelligence (eds. Liane Colonna and Stanley Greenstein)(Stockholm, The Swedish Law and Informatics Research Institute (IRI)).

[28] AI Regulation, Annex III(3)(b).

[29] AI Regulation, Annex III(3)(a).

[30] Vocational training is not defined in the law but generally it can be defined as "comprising education, training and skills development relating to a wide range of occupational fields, production, services and livelihoods." See Glossary, UNESCO, International Centre for Technical and Vocational education and Training, https://unevoc.unesco.org/home/TVETipedia+Glossary/filt=all/id=474.

[31] AI Regulation, Recital 35.

[32] *See e.g.* the case known as "Drown the Bunnies … put a Glock on their heads" where the president of Mount Saint Mary's University proposed using the results of a student survey to flag students likely to fail and urge them to drop out. As he put it, "You just have to drown the bunnies … put a Glock to their heads." *For more, see* R. Schisler and R. Golden, Mount President's Attempt to Improve Retention Rate Included Seeking Dismissal of 20–25 First- Year Students, The Mountain Echo (2016).

rather than the HE institution, relying on AI, deciding to limit access of the student to the university. It is also unclear whether Annex III's reference to "assessing students in educational training" refers to using AI to facilitate remote proctoring systems used to provide online assessments of students or whether it refers to using AI to literally assess – or score – students, through for example, some kind of grading software. Regardless, remote proctoring systems may fall under high-risk AI to the extent that they involve biometric identification, discussed more below.[33]

Where an AI system is deemed to be high-risk, then providers will have an extensive range of obligations.[34] Obligations for providers of high-risk AI systems include the adoption of risk management systems[35], data governance,[36] technical documentation[37], record-keeping[38], transparency[39], human oversight[40] and accuracy of outputs and security.[41] Additionally, providers of high-risk AI systems must put in place a quality management system.[42] Many of these requirements must be performed *ex ante* before getting access to the EU market, which will ostensibly support a legal by design approach. Users of AI systems, like universities, also have explicit obligations like monitoring the operation of the high-risk AI system on the basis of the instructions of use[43] as well as storing of logs automatically generated by the AI system.[44] Users of high-risk AI systems also need to comply with user-based rules and restrictions regarding AI system monitoring, the use of input data and the storing of logs automatically generated by the AI system. Like the General Data

[33] For a discussion on the state of the art on remote proctoring exams see forthcoming, Liane Colonna, *Legal Implications of Using AI as an Exam Invigilator*, In 2020–2021 Nordic Yearbook – Law in the Era of Artificial Intelligence (eds. Liane Colonna and Stanley Greenstein)(Stockholm, The Swedish Law and Informatics Research Institute (IRI)).

[34] See AI Regulation, Chapter II.

[35] AI Regulation, Article 9.

[36] AI Regulation, Article 10.

[37] AI Regulation, Article 11.

[38] AI Regulation, Article 12.

[39] AI Regulation, Article 13.

[40] AI Regulation, Article 14.

[41] AI Regulation, Article 15.

[42] AI Regulation, Article 17.

[43] AI Regulation, Article 29(4).

[44] AI Regulation, Article 29(5).

Protection Regulation (GDPR)[45], the proposed Regulation provides for severe penalties for non-compliance. That is regulators will be able to fine non-compliant actors up to €30m, or 6% of their worldwide turnover.[46]

The proposed Regulation provides definitions of AI "providers"[47] and AI "users"[48], referring to "public authorities" in both definitions. Here, it may be difficult to understand whether Ed Tech companies that supply products and services to HE institutions will qualify as "providers" or whether it will be the HE institutes that are given this title, and the attendant greater weight of obligations under the law. It may be particularly challenging to define public authorities as providers or users where they rely on external actors for the development of a certain AI system but put it into service under their own name. In other words, it can be very hard to distinguish between a university that makes an AI system available (through procuring an entity to build a system for it or developing it internally) ("provider") or uses an AI system ("user"). In many situations, the university is likely to be both the provider and the user. Here, the university must declare the system on the new, central database, managed by the Commission, for the registration of standalone high-risk AI systems as well as upload instructions there.[49]

As already noted, "real-time" and "post" remote biometric identification of natural persons has also been named in the proposed AI Regulation as high-risk which includes not just facial recognition, but also voice or gait recognition for identification purposes.[50] While the Commission considered a five-year moratorium on the use of such technologies in public places when initially drafting its February 2020 white paper, it ultimately decided to heavily regulate remote biometric identification sys-

[45] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) art. 4(5) (hereinafter GDPR), Article 9.

[46] AI Regulation, Article 71.

[47] AI Regulation, Article 3(2)('provider' means a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge).

[48] AI Regulation, Article (3)(4) ('user' means any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity).

[49] AI Regulation, Article 60.

[50] AI Regulation, Annex III (1)(a).

tems without going for an outright prohibition. Building on an existing legal framework with respect to biometric identification like the GDPR and the Law Enforcement Directive[51], the AI Regulation proposes to regulate all biometric identification systems yet making a controversial distinction between private and state actors as well as between uses, subjecting law enforcement's real-time use of biometric identification in publicly accessible spaces to the unacceptable risk category.

Where it concerns education, the use of biometric identification, for example, for exam invigilation or roll call[52], must comply with the high-risk systems requirements discussed above. Additionally, the compliance assessment process that is required for the producer of such a system is more stringent than the one required for any other stand-alone AI system. More specifically, the use of a system that uses AI for exam invigilation must go through a third-party conformity assessment or comply with harmonized European standards.[53] These systems will also be subject to ex-post surveillance requirements.

Finally, while the AI Regulation makes special note of the use of AI in the educational sector, it is certainly clear that not every single type of AI used in education and vocational training will be considered high-risk. For example, the use of an AI algorithm to match lecture halls and lecturers and student's course to make sure they do not clash would likely be classified as minimal or no risk. If an educational technology is classified as non-high risk, then they are not required to comply with the above requirements. Nevertheless, the provider of such a technology is encouraged to create codes of conduct.[54] Here, the idea is that the voluntary application of the above requirements would help lead to a larger uptake of trustworthy artificial intelligence in the EU.

[51] Directive (EU) 2016/680 of the European Parliament and of the Council of April 27, 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA. OJ L 119, 4.5.2016, Article 10.

[52] Carly Kind, *Containing the Canary in the AI Coalmine – the EU's Efforts to Regulate Biometrics*, Ada Lovelace Institute Blog (30 April 2021)(Explaining, "A biometric roll-call system would be classified as a high-risk biometrics system, requiring schools to ensure they procured products that had undergone a conformity assessment and received a CE marking.").

[53] AI Regulation, Article 43(1).

[54] AI Regulation, Article 69.

5 Critical perspectives

5.1 How to understand the risk categories and the appropriate level of acceptable risk?

When it comes to understanding the risk categories, many questions emerge: How does one precisely identify a system as high risk? What exactly is the process? Where does the input come from? While the AI Regulation provides a definition of high risk, categories and use cases and directs providers to take into account the likelihood and severity of the impact on health, safety, and fundamental rights there is plenty of room for ambiguity.[55] For example, what happen if low-risk AI turns into high-risk AI?

Technologies are situated in society and interact with people who can use them in ways not yet imagined. Indeed, the multistability of technologies, a concept proposed by Ihde that refers to the unpredictable uses of technology different from the originally intended ones, is well explored.[56] When it comes to Ed Tech, it is not hard to envisage that students or teachers might find unexpected uses for the technology that that neither the university nor the technology provider imagined which may be "high risk." For example, what happens when a teacher uses AI to gain insight into a student's learning habits (probably low risk), and that information, either consciously or unconsciously, impacts a student's final grade (probably high risk)? Will the software, perhaps first classified as minimal or no risk, be subjected to high-risk scrutiny by authorities considering this new application?

Another ambiguity concerns the level of acceptable risk. A provider must estimate and evaluate known and foreseeable risks that may emerge before putting an AI system on the market.[57] Even after the AI has received its CE mark and is on the market, there needs to be a continuous evaluation of risk and the provider must take measures to eliminate or

[55] *See e.g.* AI Regulation Article 6 and Article 65 (referring to Article 3(19) 'product presenting a risk' in Regulation (EU) 2019/1020 of the European Parliament and the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011.

[56] D. Ihde, *Technology and the Lifeworld. From Garden to Earth* (Bloomington and Indianapolis: Indiana University Press 1993); Mireille Hildebrandt, *Technology and the End of Law*, In: Facing the Limits of the Law (Springer 2008), 1–22.

[57] AI Regulation, Article 9.

mitigate identified risks.[58] It appears that there will always be some residual risk if the provider cannot eliminate all known and foreseeable risks and that this risk must be acceptable but who determines what level of risk is acceptable and how much discretion should it have? Also, what happens when there are unexpected benefits from the use of a technology, and should they factor into the risk analysis?

In the context of HE, understanding what constitutes a risk for health, safety, and fundamental rights, whether it can be eliminated, what adequate measures needs to be adopted to mitigate the risks and what constitutes an acceptable risk takes place in a very complex institutional setting with a distinct organizational, political, and bureaucratic culture. Is it appropriate that Ed Tech providers will largely oversee risk calculations? What happens when one Ed Tech provider or HE institution (in the event it is found to be the provider) have a larger risk appetite than another? Is it appropriate that students and teachers are not required to be consulted in the determination of what constitutes an acceptable risk?[59]

Finally, it is worth mentioning that the proposal almost exclusively focuses on risks to individuals and not risks to society. Societal risks transcend individual harm and include uses of AI systems that might harm the democratic process, the rule of law, or in this case, public education.[60] Here, a question arises concerning whether, and if so how, to consider societal risks in this already complex risk assessment. Here, a question arises whether the proposal sufficiently requires Ed Tech providers and/or HE institutions to consider the long-term societal risks and harms associated with practices like the technological surveillance of students with AI.[61]

[58] AI Regulation, Article 61.

[59] *C.f.* Article 35(9) of the GDPR (stating, "where appropriate, the controller shall seek the views of data subjects or their representatives on the intended processing, without prejudice to the protection of commercial or public interests or the security of processing operations.").

[60] Nathalie A. Smuha, Beyond the Individual: Governing AI's Societal Harm, 10 Internet Policy Review 1, 3 (2021).

[61] Alisia LoSardo, *Faceoff: The Fight for Privacy in American Public Schools in the Wake of Facial Recognition Technology*, 44 SETON HALL LEGIS. J. 373, 383–87 (2020); J. William Tucker and Amelia Vance, *School Surveillance: The Consequences for Equity and Privacy*, 2 EDUCATION LEADERS REPORT 4, 8 (2016).

5.2 Why is the focus on developers instead of universities, students, and teachers?

While most entities involved in the AI supply chain are required to comply with certain obligations, there is little doubt that AI providers are particularly burdened by the rules. Indeed, most of the requirements in the proposed AI Regulation rely on AI developers as "providers" under the regime to implement technical and organizational solutions to complex social issues. As hinted above, this may be a misstep given the complexity of the Ed Tech supply chain where many different actors are involved, including hardware and software providers, internet providers, subcontractors, etc. as well as, of course, the HE institutions, teachers and students that ultimately employ and use the tools.[62]

Where the proposed AI Regulation concerns high-risk AI in education, it asks these developers to self-assess their own compliance, at least concerning applications that do not involve biometric identification. While there are mechanisms built into the proposed regulation which encourage compliance with self-assessments like rigorous post-market surveillance, it can be questioned whether more responsibility should be placed on the HE institutions that put the systems to use, particularly as Ed Tech developers have played an increasingly prominent role in the HE in light of the COVID-19 pandemic. In other words, there is a concern that Ed-tech companies, representing private and commercial interests, may exert undue influence in the realm of public education, shifting power from the HE institutions to the providers of Ed Tech.[63] While HE institutions, as data controllers, will be responsible for the processing personal data, they may escape responsibility where, for example, anonymization techniques are applied in an application, therefore making the GDPR inapplicable.

Even where HE institutions are classified as providers and are responsible for the bulk of compliance, it still may be the case the proposal fails to sufficiently address the power imbalances between HE institutions and

[62] Forthcoming, Liane Colonna, *Implementing Data Protection by Design in the Ed Tech Context: What is the Role of Technology Providers?*, Case Western Reserve Journal of Law, Technology & the Internet (JOLTI).

[63] European Trade Union Committee for Education, ETUCE Position on the EU Regulation on Artificial Intelligence (June 2021), European Trade Union Committee for Education, https://www.csee-etuce.org/en/resources/statements/4456-etuce-position-on-the-eu-regulation-on-artificial-intelligence-june-2021:%20Liane%20Case%20Western.

students, particularly historically marginalized and under-represented students. Here, it is worth mentioning that there are no obligations for providers and/or users of high-risk AI to consult with or notify civil society organizations and affected communities.[64] The role of national authorities and standardization bodies can be juxtaposed with the role of civil society and stakeholder engagement which, is much more limited. This is regretful since it is crucial to take a relational ethics approach to algorithmic injustices and involve multiple stakeholders at the institutional or organizational levels, from the public and private sector, to nurture a dialogue on AI practices in HE.[65] The proposal's largely technocratic approach, focusing on technical fixes like data quality, fails to engage the individuals and communities that are disproportionally impacted by the AI practices.[66]

Concerning the specific role of teachers in this context, one question is whether the role of teachers is reduced to the mere providers of instructions of AI-based technologies?[67] As stated by the European Trade Union Committee for Education, "the AI Regulation should ensure that the development of AI in education does not reduce the role of teachers to mere providers of instructions but rather serves as a supporting tool for the teaching profession while preserving the professional and pedagogical autonomy and academic freedom of teachers and academics."[68] Should

[64] Forthcoming, Michael Veale and Frederik Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act,* 20 Computer Law Review International (2021)(noting, "It is unclear whether limited existing efforts to include stakeholder representation will enable the deep and meaningful engagement needed from affected communities.").

[65] Johanna Björklund, Teresa Cerratto Pargman, et. al.,WASP-HS. Community Reference Meeting: Life in the Digital World. Report (August 2021), 6–7, https://wasp-hs.org/wp-content/uploads/2021/08/WASP-HS-CRM-Virtual-Worlds-brief_Aug-2021.pdf (further explaining, "A wide range of stakeholders needs to be involved in discussing AI in higher education. Starting with students, we need to include teachers, administration, IT department, university management, trade unions, and the EdTech industry to understand better how relations constituting AI-driven educational practices are configured and shaped."); Abeba Birhane, Algorithmic Injustice: A Relational Ethics Approach, 2 Patterns 1 (2021).

[66] Abeba Birhane, Algorithmic Injustice: *A Relational Ethics Approach*, 2 Patterns 1, 2 (2021).

[67] European Trade Union Committee for Education, ETUCE Position on the EU Regulation on Artificial Intelligence (June 2021), https://www.csee-etuce.org/en/resources/statements/4456-etuce-position-on-the-eu-regulation-on-artificial-intelligence-june-2021:%20Liane%20Case%20Western.

[68] Id.

teachers have an obligation to intervene when AI gives rise to a conclusion that a student could benefit from additional support?[69]

While users of AI in HE (*e.g.* students, teachers, academics and education staff for the education sector) must be adequately informed about the intended purpose, level of accuracy, residual risks of AI tools, there is still a question about whether the AI is sufficiently transparent.[70] Will overworked academics and busy students have time to read information about the AI and, more importantly, will they have the AI literacy skills to interpret it?[71] While the proposed AI Regulation mentions the possibility of providing users with training on AI, it is unclear what this means in practice and in terms of sustainable public funding.[72]

On one hand, the proposal mainly refers to obligations for users and providers of AI systems. Here, it could be argued that there should be clearer rights for students that suffer harms because of the illegal or unethical use of AI. The proposal delegates all enforcement responsibilities to the competent authorities who can impose financial penalties and, potentially demand a noncompliant AI system to be withdrawn from the market.[73] It does not create any specific legal right to bring a claim against a provider or user for failures under the proposed law. Furthermore, the proposal does not enable an individual affected by AI practices to lodge a complaint and seek redress from a court or authority which is an especially relevant enforcement mechanism in an age where regulators have often been reluctant stand up to big technology firms (think: Max Schrems). There are no collective action mechanisms like there are in the GDPR.[74] It is also worth mentioning the proposal does not create the kinds of substantive rights for individuals such as those found in Chapter III of the GDPR ("Rights of the data subject").

[69] Teresa Cerratto Pargman, Cormac McGrath, *Mapping the Ethics of Learning Analytics in Higher Education: A Systematic Literature Review of Empirical Research*, 1 Journal of Learning Analytics 17 (2021).

[70] European Trade Union Committee for Education, ETUCE Position on the EU Regulation on Artificial Intelligence (June 2021), https://www.csee-etuce.org/en/resources/statements/4456-etuce-position-on-the-eu-regulation-on-artificial-intelligence-june-2021:%20Liane%20Case%20Western.

[71] Id.

[72] Id.

[73] AI Regulation Article 65(2), Article 71.

[74] GDPR, Article 80.

On the other hand, there exists many other laws where students can enforce their rights and complain against AI practices such as under non-discrimination laws, the GDPR, product legislation, tort law. Furthermore, the European Commission is expected to publish a draft liability framework for AI systems which could potentially strengthen the rights of individuals who are adversely impacted by AI systems. As such, it may be that the core idea behind the proposal is to enforce existing remedies rather than create new ones. This proposition is supported by reference to Article 64 which provides detailed rules for access to data and documentation by national public authorities or bodies which supervise or enforce the respect of obligations under Union law protecting fundamental rights.

5.3 When it comes to governance and oversight, who is doing what (when, and at what level?)

When it comes to governance and oversight, questions arise concerning who is doing what (when, and at what level?). The governance structure of the proposed AI Regulation involves a European as well as a national level. At the European level, with the European Commission acting as Secretariat, there exists the European Artificial Intelligence Board (EAIB), as well as the Expert Group (in planning).[75] The EAIB is tasked with collecting and sharing expertise and best practices among Member States; contributing to uniform administrative practices in the Member States; and issuing opinions, recommendations or written contributions on matters related to the implementation of the Regulation.[76]

At the national level, Member States have an important role in the application and enforcement of the proposal. Importantly, Member States must designate National Competent Authorities (NCA) to ensure the application of the law. Under the ambit of NCA is the National Supervisory Authority, Notifying Authority, and the Market Surveillance Authority (MSA).[77] The National Supervisory Authority is the authority to which a Member State assigns the responsibility for the implementation and application of the Regulation, for coordinating the activities entrusted to that Member State, for acting as the single contact point

[75] AI Regulation, Article 56–58.

[76] AI Regulation, Article 58.

[77] AI Regulation, Article 3(43).

for the Commission, and for representing the Member State at the European Artificial Intelligence Board.[78] Notifying authorities are responsible for setting up and carrying out the necessary procedures for the assessment, designation, and notification of conformity assessment bodies and for their monitoring.[79] The MSA is tasked with monitoring market activities, informing national authorities of breach of obligations, and performing activities and taking measures pursuant to Regulation (EU) 2019/1020.[80] Additionally, there are the Conformity Assessment Bodies that apply for notification and as a result become a notified body tasked with performing conformity assessments, testing, certification and inspection.[81]

These national competent authorities will have a key role for embedded AI as well as AI that relies on biometric data like Facial Recognition Technology (FRT) since these types of AI require third-party conformity assessments as well as conduct post market surveillance. It appears that the Commission would like to rely on supervisory bodies that have already been designated in accordance with other relevant Union harmonization legislation wherever possible. For example, most Member States have a body that regulates automobiles: now that body would be tasked with checking the AI in a car before giving a CE mark for the entire product. It will be harder to locate notified bodies for the new self-standing AI[82] categories like those that exist in the Ed Tech sector. Here, it appears that a Member State can opt for a sectoral approach (e.g., have its labor agency review AI tools for human resources or have its financial authority review AI tools for finance). Alternatively, it could take a more omnibus, "one stop shop" approach, delegating most tasks to a body like the Member State's DPA. It is unclear which national authority will represent each Member State in the European Artificial Intelligence Board (if the Member State takes a sectoral approach), although European Data Protection Board and European Data Protection Supervisor are already calling for

[78] AI Regulation, Article 3(42).

[79] AI Regulation, Article 3(19).

[80] AI Regulation, Article 3(26).

[81] AI Regulation, Article 33.

[82] Self-standing AI systems can be contrasted with systems that are implemented into other products like AI embedded into autonomous cars or smart toys.

DPAs to be designated as national supervisory authorities pursuant to Article 59 of the Proposal.[83]

Most Ed Tech will fall outside of high-risk AI and therefore not be subject to oversight under the Regulation. However, Ed Tech that utilizes biometric data will constitute high risk AI and therefore be subject to third party conformity assessment, at least where harmonised standards or common specifications have not been applied. It is unclear what Member State agency will be tasked with conducting the conformity assessment, but it is likely to be the DPA. Ed Tech that is used to assess students[84] or determine access to education[85] will be able to rely on conformity assessment procedure based on internal control (detailed in Annex VI), which does not require any involvement from a notified body but is nevertheless subject to post-market surveillance by the national competent authority(s).

With many different actors in the space working in different locations, times frames and with possibly different information, it is easy to imagine a lack of coordination, particularly where it concerns the monitoring of risk. This is especially true given the fact the supply chain of AI products and services is increasingly complex, distributed, and diverse.[86] The increased complexity of the supply of AI-based Ed Tech may make it harder not only for those participants acting within the supply chain to manage responsibility, as well as risk, more broadly, but also for external parties with monitoring and oversight duties.[87]

There is also a question about whether there is sufficient expertise and resources for monitoring and assessing at the national level. If expertise and resources are lacking, this may prohibit the adoption of swift qualification procedures which could have a major impact on the development

[83] European Data Protection Board, *EDPB & EDPS Call for Ban on Use of AI for Automated Recognition of Human Features in Publicly Accessible Spaces, and Some Other Uses of AI that Can Lead to Unfair Discrimination* (21 June 2021), https://edpb.europa.eu/news/news/2021/edpb-edps-call-ban-use-ai-automated-recognition-human-features-publicly-accessible_en.

[84] AI Regulation, Annex III(3)(b).

[85] AI Regulation, Annex III(3)(a).

[86] Petri Helo, Yuqiuge Hao, *Artificial Intelligence in Operations Management and Supply Chain Management: An Exploratory Case Study*, Production Planning & Control (2021).

[87] Slinger Jansen, Sjaak Brinkkemper, Anthony Finkelstein, *Providing Transparency in the Business of Software: A Modeling Technique for Software Supply Networks*, In: Establishing the Foundation of Collaborative Networks (Camarinha-Matos L.M., Afsarmanesh H., Novais P., Analide C. (eds)) (Springer, Boston, MA. 2007).

of AI within the EU. The possibly overlapping nature of these new AI regulatory bodies with DPAs may also cause confusion, and potentially undermine the authority of these bodies. Where it concerns the HE, the role of educational trade unions is unclear.[88] It is also unclear whether there will be boards similar to institutional review boards (IRB) that have an oversight role where it concerns the use of AI in HE.

5.4 Do the restrictions on biometric data go far enough?[89]

There is a question concerning whether the EU's proposal is sufficient to mitigate the potential abuse caused by technologies like FRT. There is a substantial body of research that demonstrates that the use of FRT technologies threatens marginalized communities.[90] Study after study demonstrates that FRT is typically better at detecting light-skinned people than dark-skinned people, and better at detecting men than women.[91] This, of course, raises concerns that women or students of color will disproportionately and unfairly bear the consequences of these technologies.[92] Other groups at risk for discrimination by FRT technologies include: students with accessibility needs; students with learning disabil-

[88] European Trade Union Committee for Education, ETUCE Position on the EU Regulation on Artificial Intelligence (June 2021), https://www.csee-etuce.org/en/resources/statements/4456-etuce-position-on-the-eu-regulation-on-artificial-intelligence-june-2021:%20Liane%20Case%20Western (explaining, "Education trade unions have a crucial role to play to addressing the risks of Artificial Intelligence in education and bring the perspective of AI users on the implementation of the regulation.".

[89] Part of this section is based on previous work: Forthcoming, Liane Colonna, Legal Implications of Using AI as an Exam Invigilator, In 2020–2021 Nordic Yearbook – Law in the Era of Artificial Intelligence (eds. Liane Colonna and Stanley Greenstein)(Stockholm, The Swedish Law and Informatics Research Institute (IRI)).

[90] *See e.g.*, Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, Proceedings of Machine Learning Research (2018), http://proceedings.mlr.press/v81/buolamwini18a.html (last accessed April 27, 2021).

[91] Larry Hardesty, *Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems*, MIT News (February 11, 2018), http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 (last accessed April 27, 2021); Meredith Whittaker et al., *AI Now Report 2018,* AI NOW Institute, at 16 (December 2018) *citing* Buolamwini & Gebru, *id.*

[92] Nila Bala, *The Danger of Facial Recognition in Our Children's Classrooms*, Duke L. & Tech. Rev. 249, 250–58 (2020).

ities, neurodivergence, and anxiety; low-income and rural students; and transgender students.[93]

Bias can arise both because of technical and social aspects. Technical biases arise from the way in which both hardware and software systems are designed[94] and "reduce the performance of the algorithm, hindering the achievement of its objective."[95] A core technical reason for why FRT technologies fail to identify people correctly is the use of training data sets that, for example, do not include people of African descent.[96]

An additional problem is that AI is often seen as neutral and not subject to the biases of human beings.[97] It is also the case that education is seen as neutral instead of acknowledging that is related to the changing socio-cultural and political-economic context.[98] Here, it is possible for an algorithm to be highly accurate yet be biased from a social point of view.[99] To put it differently, societal biases can be reproduced in an algorithm.[100] From an ethical and legal point of view, it can be argued that

[93] Tyler Sonnemaker, *Tech Companies Promised Schools an Easy Way to Detect Cheaters During the Pandemic. Students Responded by Demanding Schools Stop Policing Them Like Criminals in the First Place*, Insider (November 1, 2020), http://www.businessinsider.com/proctorio-silencing-critics-fueling-student-protests-against-surveilalnce-edtech-schools-2020-10?r=US&IR=T (last accessed April 27, 2021).

[94] The Institute of Technological Ethics, Three Kinds of Bias in Computer Systems, https://www.technologicalethics.org/three-kinds-of-bias (providing examples of technical bias such as "designers and programmers have a strong preference for one tool more than other tools, even though some other tools may be better or more appropriate for developing a product that will work better for achieving the purpose or end-goal as held by the product owner.")

[95] Institut Montaigne, Algorithms: Please Mind the Bias! Report March 2020, http://www.institutmontaigne.org/ressources/pdfs/publications/algorithms-please-mind-bias.pdf.

[96] Jay D. Aronson, *Computer Vision and Machine Learning for Human Rights Video Analysis: Case Studies, Possibilities, Concerns, and Limitations*, 43 Law & Soc. Inquiry 1188, 1194–95 (2018).

[97] Nila Bala, *The Danger of Facial Recognition in Our Children's Classrooms*, Duke L. & Tech. Rev. 249, 250–58 (2020).

[98] G. Biesta, *Good Education in an Age of Measurement: On the Need to Reconnect with the Question of Purpose in Education*, 21 Educational Assessment, Evaluation and Accountability 33 (2009).

[99] Institut Montaigne, Algorithms: Please Mind the Bias! Report March 2020, http://www.institutmontaigne.org/ressources/pdfs/publications/algorithms-please-mind-bias.pdf.

[100] Id.

AI should not just be "bias preserving" but also capable of improving the status quo.[101]

Where it concerns HE, universities are increasingly relying on AI-based FRT. For example, many universities have used FRT to authenticate remote users that connect from offsite the campus as well as to identify cheating and other dubious behavior throughout the online exam process during the Covid 19 pandemic.[102] The proposed AI Regulation makes a sharp distinction between identification and verification techniques, placing stricter rules on the former, and essentially placing AI used for verification purposes outside the scope of high-risk AI all together.[103] In other words, if biometric data is processed for the purpose of verification, which does not aim to uniquely identify a natural person, the processing would not fall within the categorization of high-risk AI in Annex III. While the Council of Europe has explained that biometric verification contains less risk than biometric identification because the utilization of a database is not required, it certainly contains risk such as those to fundamental rights described above.[104] Here, there is a question with the AI Regulation goes far enough to address such concerns which will no doubt be subject to great debate before the proposal becomes law.

[101] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law*, W. Va. L. Rev. (forthcoming 2021), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772.

[102] Forthcoming, Liane Colonna, *Legal Implications of Using AI as an Exam Invigilator*, In: 2020–2021 Nordic Yearbook – Law in the Era of Artificial Intelligence (eds. Liane Colonna and Stanley Greenstein)(Stockholm, The Swedish Law and Informatics Research Institute (IRI)).

[103] Id.

[104] *See* Council of Europe, Progress Report on the Application of the Principles of Convention 108 to the Collection and Processing of Biometric Data (Strasbourg: 2005), and the updated Progress Report of 2013, T-PD(2013)06, https://rm.coe.int/progress-report-on-the-application-of-the-principlesof-convention-108/1680744d81; *see also* E.J. Kindt, *Having Yes, Using No? About the New Legal Regime for Biometric Data*, 34 Computer Law and Security Review 523 (2018)(explaining, "The distinction between these two functionalities, whereby identification requires a data-base with one or more data records, is of key importance in the discussion and regulation of biometric data processing.").

5.5 Interoperability – how is this regulation going to deal with the global nature of Ed Tech?

The proposed AI Regulation is so far the most comprehensive and large-scale legislative initiative to regulate AI that has been taken and it means that the world is particularly focused on Brussels and the negotiations taking place there. Certainly, the EU sees itself as a regulatory leader where it concerns personal data protection, and it is no doubt following along the same path where it concerns the regulation of AI.[105] That said, it is unclear whether the proposed AI Regulation will have the same global impact with this proposal as the GDPR, especially given the proliferation of actors in the field already issuing soft law, hard law, and self-regulatory initiatives. It is also unclear what the proposal will mean for transatlantic AI partnerships and cooperation.

Many of the biggest AI firms are in the US and it remains to be seen whether they will be willing to adapt to the new rules to enter the EU market. While there is greater alignment in values between the democratic regimes of the US and the EU, than China, for example, and certainly some convergence where it concerns the regulation of AI (e.g., some US cities have already banned the use of FRT)[106], the US tends to prefer a more sectoral, self-regulatory approach to technology regulation with a consumer protection model of enforcement over the omnibus approach exemplified in the EU. It is likely that big tech firms that want to sell their product or service in Europe will need to adjust their systems to meet the regulatory burden, but smaller companies may simply decide to the avoid the market all together.

Broadly, the term interoperability is "the ability of diverse systems and organizations to work together."[107] The meaning of interoperability has expanded beyond its technical origins to include a diverse range of social,

[105] See Lee A.Bygrave, The 'Strasbourg Effect' on data protection in light of the 'Brussels Effect': Logic, mechanics and prospects, 40 Computer Law and Security Review 105460 (2021)(providing a detailed and critical overview of the "Brussels Effect" of EU data protection law).

[106] See Blake Montgomery, *Facial Recognition Bans: Coming Soon to a City Near You*, The Daily Beast (July 31, 2019), http://www.thedailybeast.com/facial-recognition-bans-coming-soon-to-a-city-near-you (last accessed April 27, 2021).

[107] Hunton Privacy Blog, *Interoperability: Facilitating the Global Flow of Data* (14 June 2012) https://www.huntonprivacyblog.com/2012/06/articles/interoperability-facilitating-global-flow-data/.

political, and legal frameworks.[108] In other words, while the development of AI requires interoperability of information systems (e.g., codes and architecture), it also requires the interoperability of legal systems. Interoperability can be seen as a tool to assist the growth of the digital economy, promote innovation, facilitate compliance for multinational firms and strengthen fundamental rights protections for individuals around the world.[109] Svantesson has suggested that interoperability should be a policy aim of lawmakers "to the greatest degree possible."[110]

Where it concerns Ed Tech, there is a need for both sides of the Atlantic to come together both on the development of the technology as well as to create a shared market. This is, of course, easier said than done. The 27 EU member states have a hard time creating a single market so adding the US to the mix most certainly adds complexity. On one hand, making transnational entities choose between conflicting regulatory frameworks is regretful at a time when promoting legal interoperability is critical to support the development of AI applications in key sectors like education. On the other hand, it is possible, in the long run, that the proposal will lead to interoperability through the creation of a common legal, ethical, and technical standards.

6 Conclusion

This paper has made some preliminary observations as well as offered some critical perspectives concerning the way the proposed AI Regulation addresses the educational context, particularly HE. It concludes that the proposal represents a fundamental shift in approach towards the more ethical use of AI in the HE, albeit one that suffers from certain defects

[108] *What is Interoperability?*, Network Centric Operations Industry Consortium, https://www.ncoic.org/technology/what_is_interoperability.

[109] Hunton Privacy Blog, *Interoperability: Facilitating the Global Flow of Data* (14 June 2012) https://www.huntonprivacyblog.com/2012/06/articles/interoperability-facilitating-global-flow-data/.

[110] Dan Jerker B Svantesson, *The Holy Trinity of Legal Fictions Undermining the Application of Law to the Global Internet*, 23 International Journal of Law and Information Technology 219, 234 (2015)(stating, "(o)ur aim should be to create jurisdictional interoperability between the different domestic legal systems to the greatest degree possible… by identifying any uniting features (of which there are many), and in seeking to iron out inconsistencies and clashes, between domestic legal systems, both in substantive and procedural rules, much can be achieved.").

that may undermine its ultimate effectiveness as a mechanism to ensure accountable, transparent, and responsible AI. These defects include the difficulty of understanding high-risk applications of AI in the Ed Tech sector as well as a lack of focus on universities, students, and teachers who ultimately employ and use the tools. When it comes to governance and oversight, there are a multiplicity of actors at both the national and EU level that possess different competences, interests, and capabilities. This introduces complexity and possibly a lack of coordination that may undermine effective governance. There is also a question about whether there is sufficient expertise and resources for monitoring and assessment at the national level. Furthermore, there is an issue concerning whether the EU's proposal is sufficient to mitigate the potential abuse caused by technologies like FRT in the HE context. Finally, it is unclear how this regulation will deal with the global nature of AI-based Ed Tech and promote legal interoperability in the realm of AI.

Cecilia Magnusson Sjöberg & Rebecka Weegar

Means for Memo Matching (MMM): A Study of Legal Informatics and Language Technology

1 Project approach

1.1 Project team

This chapter is about the Means for Memo Matching (MMM) Project and how it has enabled studies of legal informatics[1] and natural language processing[2] in higher education[3]. Artificial intelligence (AI) tools have been one attribute for promising results. The research has increasingly been carried out over the last couple of years on an ad hoc basis at two Stockholm university departments, namely the Law Department and the Department of Computer and System Sciences.[4] It should be empha-

[1] *Legal informatics* is commonly understood as a technologically oriented intersection of the research field Law & ICT (information and communication technology). For more on that kind of approach, see *Legal Management of Information Systems – Incorporating Law in e-Solutions*, Cecilia Magnusson Sjöberg (ed) (Studentlitteratur 2005). The other field within this context is usually labelled (substantive) *ICT Law* and takes an interest in how to interpret and apply law in digital environments, such as the internet. See further, e.g., *Rättsinformatik i det digitala informationssamhället*, Cecilia Magnusson Sjöberg (ed) (Studentlitteratur 2021). See also Cecilia Magnusson Sjöberg, 'Legal Automation: AI in Law revisited' in Marcelo Corrales, Mark Fenwick and Helena Haapio (eds), *Legal Tech, Smart Contracts and Blockchain* (Springer 2019) pp. 173–187.

[2] *Natural language processing* is an area of research and a set of methods and technologies for processing human language with computers.

[3] This refers primarily to university education.

[4] List of participants: *Cecilia Magnusson Sjöberg*, *Stockholm University*, LL.D., Professor of Law & Informatics, Subject director, *Rebecka Weegar, Stockholm University* PhD, Lec-

sised that the text here presented is merely a beginning of forthcoming research in this environment. More precisely the notion of MMM works as a trigger of investigations into the interplay of various kinds of matching of legal texts such as machine grading versus manual grading etc. So, in this project legal and computer science researchers collaborate on the question if the grading of a short-written assignment in higher education can be fully or partly automated with the use of AI (artificial intelligence) tools. The legal researchers in the project have an approach based in legal informatics. The computer scientists mainly draw on expertise from the field of language technology.

A good project team is essential in many aspects. The current MMM Project is an example thereof. In this context, the research requires an understanding of the interplay between law, language, and technology. In the MMM Project, emphasis is mainly placed on methodological issues, but knowledge of facts and other substantive matters is also taken into consideration. Examples of substantive issues include basic information about the normative hierarchy of legal sources such as constitutional laws and (decided) court cases as well as linguistic classification systems. In other words, *a mix of skills* is needed in a project of this kind, and these skills must in its turn be inserted and integrated into the analysis. For instance, it can be noted that the MMM Project team includes both junior and senior researchers.

1.2 Starting points

One initial and major assumption in the MMM Project is that grading at universities can under certain circumstances be performed wholly or partially *automatically*. This implies that full automation is not a goal. A second assumption is that the generic and multifaceted notion of *grading* needs to be specified. Thirdly, we assume that *AI-based* solutions are promising in learning analytics.

The *setting* of this study has as mentioned above been the Law programme at the Department of Law, Stockholm University, in collaboration with the Department of Computer and Systems Sciences (DSV). The test material is a compulsory short written assignment (one page memo, see Annex 1 for more details on the writing instructions) in which

turer at Department of Computer and System Sciences, *Johan Rosell, Stockholm University*, Research assistant.

students have to discuss the General Data Protection Regulation[5] from a methodological point of view. To be a bit more precise major grading features are (a) *facts*, i.e. how well a student is able to relate to adequate data in the current situation. Next step is (b) *focus*, i.e. the ability to apply an analytical approach. Finally (c) *form* is relevant, i.e. professional document management.

The students primarily concerned are those taking today's mandatory course 'Rättsinformatik' (Legal informatics) during the fourth year of the Law programme at Stockholm University. More information on relevant parts of the syllabus, etc., will follow below. It is important however, already here to note that the course in question reflects also in general terms how Law & information and Communication Technology (ICT), since the beginning of the 1980s, has played an important role at Stockholm University, not only within legal education but also in teaching for instance tech students.[6] Other components currently include digitalisation and internationalisation in the light of privacy and data protection, automatic and autonomous decision-making and legal aspects of information security. Letting law play a *proactive role*, instead of merely functioning as a reactive conflict-solving mechanism when things have already gone wrong, is a critical success factor.

Common denominators within the MMM Project are *grading* and *graders*. In this context it is important to note that the grading of the mandatory task of completing a written assignment – a methodologically oriented memo – on the topic *General Data Protection Regulation* is not equivalent to the more differential grading scale used in the students' final course grades. Instead, students receive feedback in the form 'fail', 'good' or 'very good'. In order to receive a final course grade, the requirement is a passing grade ('good' or 'very good'). The more fine-grained categorisation is made so as to reflect the structure of the final *exam*. When it comes to graders, there are a variety of set-ups. Graders can be more or less qualified, interested in the topic area, pedagogically skilled, etc. In the MMM Project, we included one senior grader and one junior grader. To conclude, there is a major distinction to be made between

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation), commonly referred to as the GDPR.

[6] For such an approach see, e.g., *IT Law for IT Professionals – an Introduction*, Cecilia Magnusson Sjöberg (ed) (Studentlitteratur 2005).

grading of the one-page written assignment on the one hand and the final general course grade on the other. As long as a student pass the written assignment the result will not have any particular impact on the final course grade as such which is given on the scale “AB”, “BA”, “B”, or “underkänd”/“fail”.

A project of this kind can easily become comprehensive when it comes to data collection, processing, and management. For an overview, see parts of the Memo corpus in Annex (B).[7] For reasons that we will describe below, we chose to delimit the primary scope to only ‘fail’ in certain parts of the study. A consequence of this delimitation was a need for ‘fail’ features that became a task in itself within the project.

1.3 Structure of contents

The contents of this chapter are structured in the following way. The text begins with the *project approach* described in terms of how the project team was composed as regards scientific skills and seniority. Points of departure are then conventionally expressed as hypotheses to be verified or falsified. The part that reports on *project activities* is vital for the study. From a legal point of view, an overview of the *legal framework* is also crucial. The concluding remarks will probably be of greatest interest to the reader. In addition, there is some documentation to be found in the *annexes*. *References* are naturally listed in footnotes.

2 Project activities

2.1 Research set-up and the data used in the study

Given the fact that this book chapter is quite atypical contributing to the legal domain in a digital setting rather than in the traditional theoretical dogmatic format. This is the overall explanation why conditions and outcome of the studies are presented in a seemingly abstract format way of notes rather than full sentences in the traditional way. General aspects of the included studies are thus listed as a next step.

[7] On this kind of research, see for instance Cecilia Magnusson Sjöberg, *Critical Factors in Legal Document Management: A Study of Standardised Markup Languages. The Corpus Legis Project* (Jure 1998).

Included MMM studies – general aspects

Memos were generated each semester (approx. 200–250)
Project Manager: study outline (design)

Manual grading Grader 1 and Grader 2
Pilot study 1A Memos 1–24
Pilot study 1B Memos 25–49
These memos were graded separately and then jointly

Manual grading Grader 1 and Grader 2
Major study 2 Memos 50–499
Grader 1 Memo 50–274
Grader 2 Memo 275–499
Graded separately
Grades were not negotiated in the major study.

In summary, when using the machine with training and validation data, the researchers used:
– 432 (450 - 18) memos in the range 50–499.
– Five extra memos graded 'fail' were added (to increase the number of memos in that particular dataset).
– the 49 memos from the pilot study.

For the purpose of use as test data, the grading of 18 memos (the ones divisible by 25) were extracted, see above.

Pilot studies (1)

1A Memos 1–24 (= 24)
All 24 memos were first graded separately by two graders (without any prior discussion).
Grades were negotiated for the purpose of future consistency in grading.

1B Memos 25–49 (= 25)
All 25 memos were graded separately (there was some synchronization between the two graders from pilot study 1A).
Grades were negotiated for the purpose of future consistency in grading.

Major study (2)

2

Memos 50–499 (= 450) plus 5 additional memos graded 'fail' and the 49 memos from the pilot study, i.e., in total 450 + 5 + 49 = 505 memos Multiple applications were run.
274 were graded by Grader 1 and memos 275–499 were graded by Grader 2. Note that there was some synchronization between the two graders from pilot studies 1A and 1B.

– 49 (randomly selected) memos from within the range 275–499 were also graded by Grader 1, to be used for consistency between the two graders in the major study. These 49 memos were used as the validation dataset.

– The results of the grading of 18 memos (every 25th in the range 50–499) were retained by the graders for use as test data in the comparison between 'man' (grader) and 'machine' (algorithm) as a final test dataset. The other results of the grading (432 memos) were used as training and validation data. As mentioned, there was some synchronization between graders from pilot studies 1A and 1B.

The following observations appear to be of particular interest within the MMM Project. To begin with, *the end result* including the classifier is in itself interesting. The *classifier* can simplified be described as the machine generated classification "model", based on training data, for sorting written assignments into the categories 'fail' or 'pass' respectively. We have also noted a co-existing consistency as well as discrepancy among *junior and senior graders*. This applies also internally with regard to each *individual grader's* consistency with himself/herself. Furthermore, there are potentials associated with a *combinational approach* (human beings and AI: training data, validation data and test data). Mention should also be made of the impact of *negotiations* among graders), e.g. in terms of unwanted vagueness.

2.2 Narrowing down the scope of the analysis

One task that emerged during the study was a need to limit the scope of the analysis to only two output categories[8] (fail/pass) instead of three (fail/good/very good). From the start, there was a relatively clear distinction between 'fail' on the one hand and 'good'/'very good' on the other for both the human graders and the machine learning (ML) classifier. However, the distinctions within the category 'good/very good' were much vaguer. Therefore, the decision was made to focus on *the 'fail' assessments and underlying 'fail' features*. This has surely had some impact on the end results, but the authors believe the delimitation was justified.

To illustrate the concept, a few 'fail' features identified by the graders and later used by the ML classifier are listed below. The left-hand side shows what might be referred to as features that should be included in a memo, and on the right-hand side, a few features that should be excluded from a memo are mentioned.

'Fail' features

Should be included in text

- Sufficient number of words
  - Comprehensiveness
- Paragraphs
  - Readability
- Editing language
  - English
- References
  - Articles
  - Governing frameworks
- Important concepts
  - Controller
  - Data subject
  - GDPR
  - Privacy

Should be excluded from text

# Checklists
# Grammatical mistakes
# Plagiarism

[8] In section 4 these are called "labels". The terms output categories and "labels" can be used interchangeably.

3 Legal framework

The legal framework surrounding the project primarily in terms of applicable rules and regulations consists of a multitude of smaller parts. Here it is important to emphasize that the MMM Project is more or less completely methodologically oriented. The next stage in the work will however broaden the analysis (scope) towards substantive (material) law. The overall ambition has therefore been to review and ensure legal compliance, rather than to perform in-depth analyses of the law in force. Such exercises can already be found in the legal doctrine addressing the legal implications of information and communication technologies. Instead, at this stage of the MMM Project primary concern has been that personal data was processed in accordance with the General Data Protection Regulation, as there is no doubt that the MMM Project involves *personal data* processing that falls within the scope of the GDPR.

The kind of provisions that need to be taken into consideration can be exemplified by governing legal definitions (Article 4) and the important distinction between anonymisation (where the General Data Protection Regulation does not apply) and pseudonymisation (where the General Data Protection Regulation does apply). Further, there are general data protection principles (Article 5) that must be adhered to, such as lawfulness, fairness, and transparency. Of utmost importance is the requirement that the so-called controller has a legal ground for the processing, e.g., the data subject's consent, making processing lawful (Article 6). The information duties (Articles 12–15) can in practice be quite burdensome, as they comprise both information to be provided upon the initiative of the controller and also upon the request of the data subject (Articles 12–15). Applied automated decisions, including profiling, are another aspect of the algorithms and associated models in the project (Article 22). Attention should also be paid to legal system development (Article 25), i.e., data protection by design and by default. For further reflections concerning for instance fulfilment of information duties, see the annexes.

Another regulation of great importance is the European Commission's proposal COM(2021) 206 final, for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending certain Union legislative acts. As an example, mention could be made of the 44 legal definitions laid down in Article 3 of the proposal, comprising for instance the following concepts relevant to artificial intelligence (AI):

> (29) 'training data' means data used for training an AI system through fitting its learnable parameters, including the weights of a neural network;
>
> (30) 'validation data' means data used for providing an evaluation of the trained AI system and for tuning its non-learnable parameters and its learning process, among other things, in order to prevent overfitting; whereas the validation dataset can be a separate dataset or part of the training dataset, either as a fixed or variable split;
>
> (31) 'testing data' means data used for providing an independent evaluation of the trained and validated AI system in order to confirm the expected performance of that system before its placing on the market or putting into service;

In addition to data protection regulation and a compliance check with the proposed AI regulation, there is quite a lot of legislation that one must be aware of and comply with when the setting is the *public sector*. This is the case in the MMM Project: in Sweden publicly funded universities are government agencies (*myndigheter*). This means that the automated grading in the MMM project should be compliant with both general and special *administrative law* governing teaching activities, including examination measures of different kinds (such as those included in the MMM Project). Fundamental principles of *openness* capturing transparency[9] vs. *secrecy* (confidentiality) are also to be considered during system design, development, implementation, and management. From a legal point of view, legal *digital archives* that are synchronised with daily information flows are also on the regulatory wish list.

Last, but not the least, *ethical considerations* must be made. It is important to let ethical vetting and similar measures play a separate role, so as not to be directly incorporated as law (generally speaking).

4 Machine learning approach

The topic of the MMM study is the automatic scoring of written assignments or essays[10], which falls within the research area of natural language processing. Natural language processing is a subfield of artificial intelli-

[9] Regarding this interplay, see Cecilia Magnusson Sjöberg, 'Legal AI from a Privacy Point of View: Data Protection and Transparency in Focus' in Sonya Petersson (ed), *Digital Human Sciences* (Stockholm University Press 2021) DOI: https://doi.org/10.16993/bbk.h.

[10] AES is an acronym for automated essay scoring.

gence that concerns automatic processing of spoken and written human language.

Often, some kind of machine learning is used for automatic essay scoring[11]. Machine learning (ML), another subfield of AI, can be used as a tool to accomplish a wide array of tasks involving data, and the goal is often to train a classifier to mimic the behaviour of an expert in some field. This can be accomplished by applying a learning algorithm to examples labelled by an expert. In the case described here, the expert is the grader, the task is assessment of student texts, and the data are texts written by students, with associated feedback labels: 'pass' or 'fail'. The result of the training is a classifier adapted to assessing student texts that are similar, but not identical (this would be plagiarism, for which there are other tools for discovery), to those in the training data. In this case, the classifier is adapted for the particular student assignment described in this study.

4.1 Text features

When applying machine learning to assessment of students' assignments, it is important to consider how to represent the students' texts. One main component is of course the contents of each text: the words that the student has used. However, there are also other characteristics of a text that can influence the assessment or that can indicate the overall quality of the text. Examples include the length of the text, the vocabulary, and any errors in spelling or grammar. The characteristics selected to represent a text are called *features* and the selection of features has a large impact on how well a machine learning classifier can be trained to perform classification.

In this study, a number of different features have been included. They include the 'features of fail' discussed in section 2.2, and also features that describe the structure of a text, such as the number of paragraphs and headings therein[12] (see further Annex C).

[11] For a survey on notable AES systems, see: Semire Dikli, 'An Overview of Automated Scoring of Essays' (2006) The Journal of Technology, Learning and Assessment 5.1. For an overview of recent AES research, see Ke, Zixuan, and Vincent Ng, 'Automated Essay Scoring: A Survey of the State of the Art' (2019) IJCAI vol. 19.

[12] Three libraries were used to generate features, pyspellchecker by Peter Norvig (https://norvig.com/spell-correct.html), language-tool-python (https://pypi.org/project/language-tool-python/), and Natural Language Toolkit, see Steven Bird, Edward Loper and Ewan Klein, *Natural Language Processing with Python* (O'Reilly Media Inc. 2009).

4.2 Outlier features

A majority of the students' texts were similar in form and content, but some texts had more unusual characteristics. We therefore included features to identify texts that were significantly different from the 'average' text. For example, a majority of the texts consisted of around five paragraphs, but a smaller number of texts consisted of either one long paragraph or many, very short paragraphs, resembling a list. If the number of paragraphs deviated significantly from the norm, information about this was included as a binary feature[13]. Other examples of features in this category were unusually few references to articles in the General Data Protection Regulation and a text being unusually short. The total number of outlier features for each text was also included as a feature.

These outliers were used to automatically generate feedback comments which could be displayed either to a grader or directly to the student who had handed in the text.

A feature that is useful in machine learning should correlate with the target label, in this case the 'pass'/'fail' assessment. The figure in Annex C shows the correlation between a selection of the included features and the 'fail' label. In that figure, it can be noted that the 'outlier features' have a stronger correlation with the feedback label ('pass'/'fail') than the original features based on the number of occurrences of some characteristic. For example, knowing the number of paragraphs is less informative than knowing that the number of paragraphs is either very low or very high compared with the average for all texts.

4.3 Training the classifier

To develop a classifier based on machine learning, a suitable dataset must be selected. As mentioned in the examples of AI relevant concepts in section 3, the dataset is usually divided into distinct parts. In this study, a majority of the data were used for training and the remaining data were used for validation and evaluation of the trained classifier.

A common strategy is to use a validation dataset during development, for instance to select which features to include. An alternative (or complementary) approach is to apply cross-validation. During cross-validation, the training data are further divided into smaller segments. A

[13] A binary feature represents a characteristic that is either present in the text or not.

model is trained on all but one of the segments and then evaluated on the remaining data. This is repeated until all segments have been used for evaluation. This method is suitable when limited training data are available, since it eliminates the need for a separate validation dataset. Here, cross-validation was used to select and finetune the learning algorithms, and to select which features to include. The best choices found during cross-validation were evaluated against the validation data and the final test dataset.

In all, 18 memo texts were set aside as a test dataset, 49 memo texts were used for validation and the remaining 437 texts were used to train the classifier. For the training data stemming from the pilot study, the negotiated grades ('feedback') of the two human graders were used as labels. The remaining training data was only graded by one of the two graders meaning that no negotiation was employed for labelling. The labels, 'pass' and 'fail', assigned by the graders were thus used as the 'ground truth', where the goal of training is for the classifier to be able to assign these labels in a way that mimics the human labelling.

4.4 Evaluation and performance measures

Two important measures for evaluating the performance of a classifier are recall and precision. When the task is to classify texts into assessment categories, a high precision value for a category like 'fail' means that the texts assigned to that specific category truly belong in that category. A high recall means that the classifier correctly, and among all available texts, identifies the texts that should belong in a specific category. Precision and recall values can range from 0 to 1, where 1 is a perfect score. A recall of 1 for a category means that the classifier has correctly identified all the texts in that category, while a recall of 0.5 means that the classifier has missed half of the texts.

A good classifier should have high scores for both precision and recall, but there is often a trade-off: when you increase precision, you might decrease recall and vice-versa. Imagine that we have a classifier that classifies all assignments as 'fail'. This corresponds to a recall of 1 for the label 'fail': every text that deserves the feedback 'fail' will be labelled as 'fail', but the precision would be low, as we would fail many students that deserve to pass. On the other hand, if the classifier would only assign the feedback 'fail' to a single text deserving of that feedback, the classifier would have perfect precision for the label fail, since no students that should pass would be failed. Yet, this would also result in many false negatives (students that should have failed get a pass), meaning that recall would be very low. While the ideal is that a classifier has both high precision and recall, one of the measures can be prioritized over the other.

In this case, we were particularly interested in a high recall for the category 'fail', as one goal was to correctly identify all the texts that lacked some quality necessary to get a passing grade on subsequent assignments in the course.

4.5 Agreement and Cohen's kappa

Ideally, a classifier that has been trained for a specific task should agree with human experts performing the same task. However, how well a classifier can be expected to perform varies depending on the type and complexity of the task. One way of estimating an upper bound for the expected performance is to measure the agreement between two or more experts on the same task. A well-defined task should yield a high level of agreement, while a more complex or less well-defined task can result in lower levels of agreement. A common way of measuring agreement between experts in text classification and automatic grading and assessment is to use Cohen's kappa, κ[14]. The maximum possible value for this measurement is 1.0. Here, the texts in the validation data were graded by two graders, independently, and the κ for them was calculated to be 0.43[15], which indicates a certain level of agreement, but not complete agreement. One of the graders assigned 'fail' as feedback to five texts in the validation dataset, and the other grader assigned 'fail' as feedback to seven texts. In all, three texts were given the feedback 'fail' by both graders.

[14] Cohen's kappa, κ, measures agreement between two assessments while adjusting for chance agreement, see Jacob Cohen, 'A coefficient of agreement for nominal scales' (1960) Educational and Psychological Measurement 20.1, 37–46.

[15] For an interpretation of kappa scores, see J. Richard Landis and Gary G. Koch, 'The measurement of observer agreement for categorical data' (1977) Biometrics, 159–174. They denote a score between 0.41 and 0.6 as moderate agreement, and require a score of at least 0.61 for substantial agreement.

4.6 Partially automated grading

There were two main challenges in constructing a useful classifier for the task described here. First, 'fail' was a *minority class*, meaning that there were only a handful of texts (13%) that received a 'fail' assessment in the training data. Therefore, the learning algorithm had only a few instances to learn from, making it difficult for it to determine which feature patterns corresponded to a likely assessment of 'fail'. Second, the manual grading of the validation dataset showed a fairly low agreement between the two graders. This means that some texts with similar characteristics were possibly assigned contradictory labels in the training data.

For these two reasons, it was not expected that a classifier trained on these data could fully replace an expert grader, which caused us to set a second goal: to reduce the number of texts needing manual grading by half. Such 'partially automated grading' could be achieved by letting the classifier divide the texts into two groups: one group of texts with a high probability of a passing grade and one group of texts with some probability of a failing grade. Manual grading would only be needed for the group with some probability of 'fail', while all other texts could automatically be given a passing grade.

The classifiers were therefore used to rank all the texts in the test and validation datasets, from highest to lowest probability of 'fail', the goal being that all texts in the category 'fail' should end up in the top half of the list, while the texts in the bottom half could be considered as passing. This would be possible for a classifier that assigns a probability for each text to belong to the class *Fail*. Usually, if this probability exceeds 0.5, the class *Fail* would be assigned by the classifier. Here, we lowered the probability threshold until half of the texts were placed in the 'possible fail' group. This can also be understood as increasing the threshold for considering a text as belonging to the class *Pass*. Only texts with a very high probability of belonging to the class *Pass* would be assigned to that class by the classifier.

4.7 Results of the automatic text assessment

Two different learning algorithms performed well during the cross-validation: Random Forest (RF) and Gaussian Naive Bayes (GNB)[16]. These final classifiers were evaluated against both the validation dataset and the test dataset. The validation dataset consisted of 49 texts that had been graded by two graders independently and the performance of the classification model was evaluated against both expert graders for both the RF and the GNB classifiers. Cohen's kappa values for the validation dataset with both classifiers and both graders are shown in the table below:

Evaluation set:	κ, Random Forest	κ, Gaussian Naive Bayes
Validation set assessed by Grader 1 (49 instances)	0.56	0.76
Validation set assessed by Grader 2 (49 instances)	0.46	0.56
The 43 instances in the validation set given the same feedback by both graders	0.79	0.84

These numbers can be compared to the agreement value between Grader 1 and Grader 2 for the validation dataset, which was 0.43. In terms of κ, both classifiers agreed with each of the graders individually to a greater degree than the graders agreed with each other.

When evaluating the classifiers on the subset of the validation dataset where Grader 1 and Grader 2 had assigned the same grade (the last line in the table), GNB achieved the highest agreement, 0.84. This dataset could be considered as containing student texts which are 'easier' to grade, since the borderline cases where the two graders disagreed were removed. Still, this subset could also be considered a more reliable evaluation dataset, as there were no disagreements regarding these texts.

Comparing the two classifiers in terms of precision and recall (here, both are compared with Grader 2), the RF classifier had higher precision overall, while the GNB classifier had higher recall overall. This means, that while all texts that were classified as 'fail' by the Random Forest

[16] Implementation from the Scikit-learn library: Pedregosa *et al.*, 'Machine Learning in Python' (2011) Journal of Machine Learning Research 12, 2825–2830.

classifier, were also labelled as 'fail' by Grader 2, less than half of the texts that should have been classified as 'fail' were identified. The higher recall of the GNB classifier, on the other hand, meant that the GNB classifier was more suitable for identifying the texts that should be classified as 'fail' which corresponds with our goal: to identify all the texts that should possibly be assessed as 'fail':

	Precision	Recall
Random Forest, Fail	1.00	0.43
Random Forest, Pass	0.91	1.00
Gaussian Naive Bayes, Fail	0.75	0.86
Gaussian Naive Bayes, Pass	0.98	0.95

For the final test dataset, consisting of 18 student texts graded by Grader 1, both classifiers were applied again. This dataset had not been available during the development of the classifiers and contained only one text assessed as 'fail'. When applying the classifiers to this dataset, the RF classifier did not manage to correctly identify that text. However, it was identified by the GNB classifier, which achieved a precision of 0.33 and a recall of 1.0 for the '*fail*' class and a precision of 1.0 and a recall of 0.88 for the '*pass*' class.

The GNB classifier assigning the label 'fail' to three of the texts in the test set, with the text manually labelled as 'fail' being among those three texts. If we could trust the classifier to produce the same quality of results in the future, it would be possible to manually review only the texts classified as 'fail' by the GNB classifier, while automatically passing the remaining texts. This would lead to a large reduction of manual work in grading, as about 80% of the texts would be automatically labelled.

However, for the partially automated grading, discussed in section 4.6, a threshold of 50% was set, where half of the texts would be automatically labelled as 'pass', while the other half would be given manual feedback. This approach requires more manual work, but reduces the risk of incorrectly labeling a text with 'pass'. For this partially automated grading, both classifiers managed to correctly divide the texts in the validation dataset and in the test dataset so that only 50% would require manual grading, with no 'fail' texts being missed. This corresponds to a 'fail' recall of 1.0 for both classifiers and both datasets.

5 Concluding remarks

The MMM Project could briefly be described as a legally oriented text analysis of grading assessments in higher education (HE). The text analysis is based on language technology used for the purpose of classification by means of machine learning. More precisely, the project is about means for memo matching and enhanced equal treatment of students by automation, combined with individual manual feedback. The combination of *man (humans)* and *AI (here machine* learning) proved to be important.[17]

The overall goal was to achieve accurate and efficient grading in a specific exam situation that could involve one or several grader(s). *Consistency* was considered essential. Mention could here be made of a very limited test in the pilot study that resulted in four consistent, manually self-graded memos. Ultimately, there are two major categories of consistent grading (Grader 1 to Grader 2, Grader X to machine) and one category of consistent self-grading (Grader X to Grader X). (Here "Grader X" stands for a grader that has not been specified as either Grader 1 or Grader 2, the overall purpose being remaining anonymity.)

Along with the interplay of man and machine and issues of consistency comes the role of *negotiations* within a framework that allows *consultation* among graders. *Diversified manual grading* is another result of the study. This implies that there is a considerable variety among human graders, which in turn opens for unwanted vagueness and limitations in foreseeability. A rather complicated outcome is when two seemingly qualified graders make strikingly diverging assessments. For instance, in this study, there was one memo assessed as 'fail' by one grader and as 'very good' by another.

In a follow-up analysis, there were some indications of *seniority* among graders as a critical factor for giving a passing grade to 'odd', but accept-

[17] In total, 450 memos were included in the project. Initially, the assumption was that half of the included memos would be graded completely automatically. Furthermore, it was assumed that half of the included memos would require supplementary manual assessment (based on a digital selection). However, conditions changed throughout the pilot study (see above).

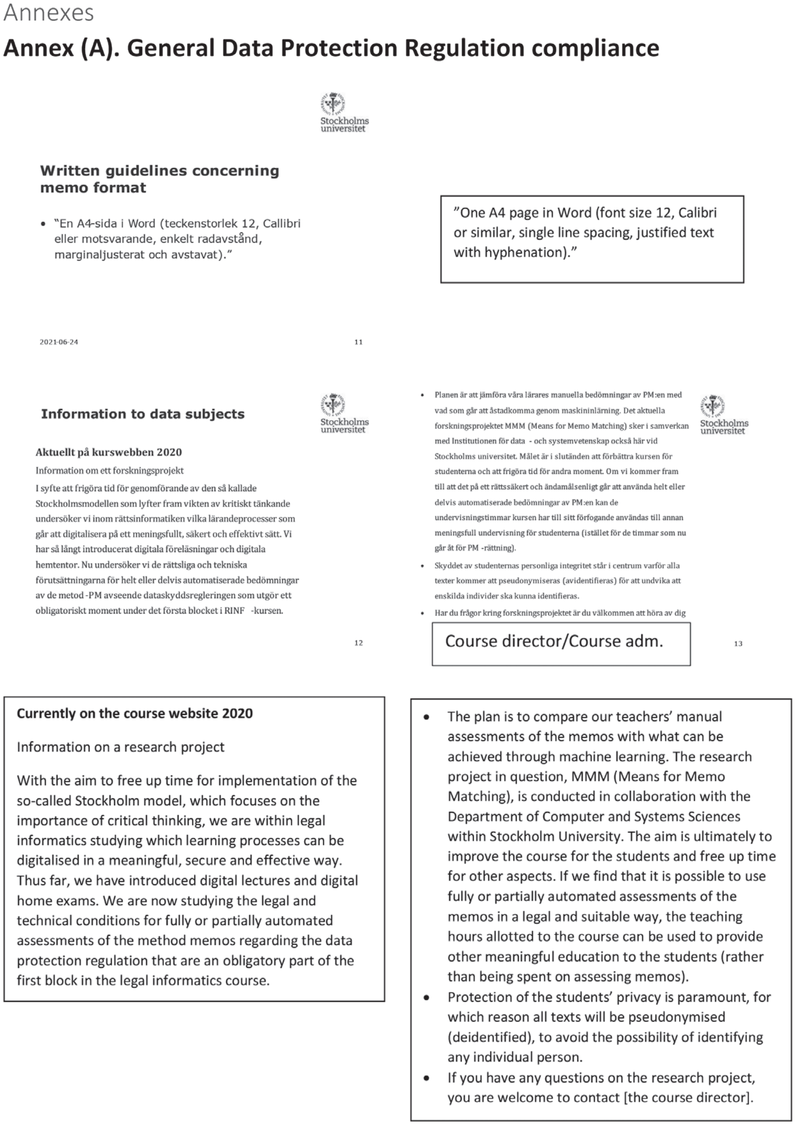
able written assignments. With an outlook focusing on *learning analytics* and achieving progress within the project, the decision was made to narrow down the scope of written assignments included. More precisely, the perspective was shifted from a general approach to *fail assessments and features*.

From an automatic grading perspective, the main result was that the hypothesis that it would be possible to reduce manual grading by at least 50% was confirmed with both the validation dataset and the test dataset using two different classifiers.

The outlier features were found to be particularly useful for classification. For example, while a student text in this case should contain content from the General Data Protection Regulation, too much overlap with the GDPR could indicate that the student might not have added much content of their own. With a larger dataset, the classifier could be expected to identify such a large-but-not-too-large overlap pattern. However, with only few examples of this particular pattern, the simple fact that a text diverges from the norm can be highly informative for classification. This approach could be further developed for automatic grading in general.

Annexes

Annex (A). General Data Protection Regulation compliance



Annexes

Annex (A). General Data Protection Regulation compliance

Stockholms universitet

Written guidelines concerning memo format

- "En A4-sida i Word (teckenstorlek 12, Callibri eller motsvarande, enkelt radavstånd, marginaljusterat och avstavat)."

"One A4 page in Word (font size 12, Calibri or similar, single line spacing, justified text with hyphenation)."

2021-06-24 11

Information to data subjects

Stockholms universitet

Aktuellt på kurswebben 2020

Information om ett forskningsprojekt

I syfte att frigöra tid för genomförande av den så kallade Stockholmsmodellen som lyfter fram vikten av kritiskt tänkande undersöker vi inom rättsinformatiken vilka lärandeprocesser som går att digitalisera på ett meningsfullt, säkert och effektivt sätt. Vi har så långt introducerat digitala föreläsningar och digitala hemtentor. Nu undersöker vi de rättsliga och tekniska förutsättningarna för helt eller delvis automatiserade bedömningar av de metod -PM avseende dataskyddsregleringen som utgör ett obligatoriskt moment under det första blocket i RINF -kursen.

12

- Planen är att jämföra våra lärares manuella bedömningar av PM:en med vad som går att åstadkomma genom maskininlärning. Det aktuella forskningsprojektet MMM (Means for Memo Matching) sker i samverkan med Institutionen för data - och systemvetenskap också här vid Stockholms universitet. Målet är i slutänden att förbättra kursen för studenterna och att frigöra tid för andra moment. Om vi kommer fram till att det på ett rättssäkert och ändamålsenligt går att använda helt eller delvis automatiserade bedömningar av PM:en kan de undervisningstimmar kursen har till sitt förfogande användas till annan meningsfull undervisning för studenterna (istället för de timmar som nu går åt för PM -rättning).
- Skyddet av studenternas personliga integritet står i centrum varför alla texter kommer att pseudonymiseras (avidentifieras) för att undvika att enskilda individer ska kunna identifieras.
- Har du frågor kring forskningsprojektet är du välkommen att höra av dig

Course director/Course adm.

Stockholms universitet

13

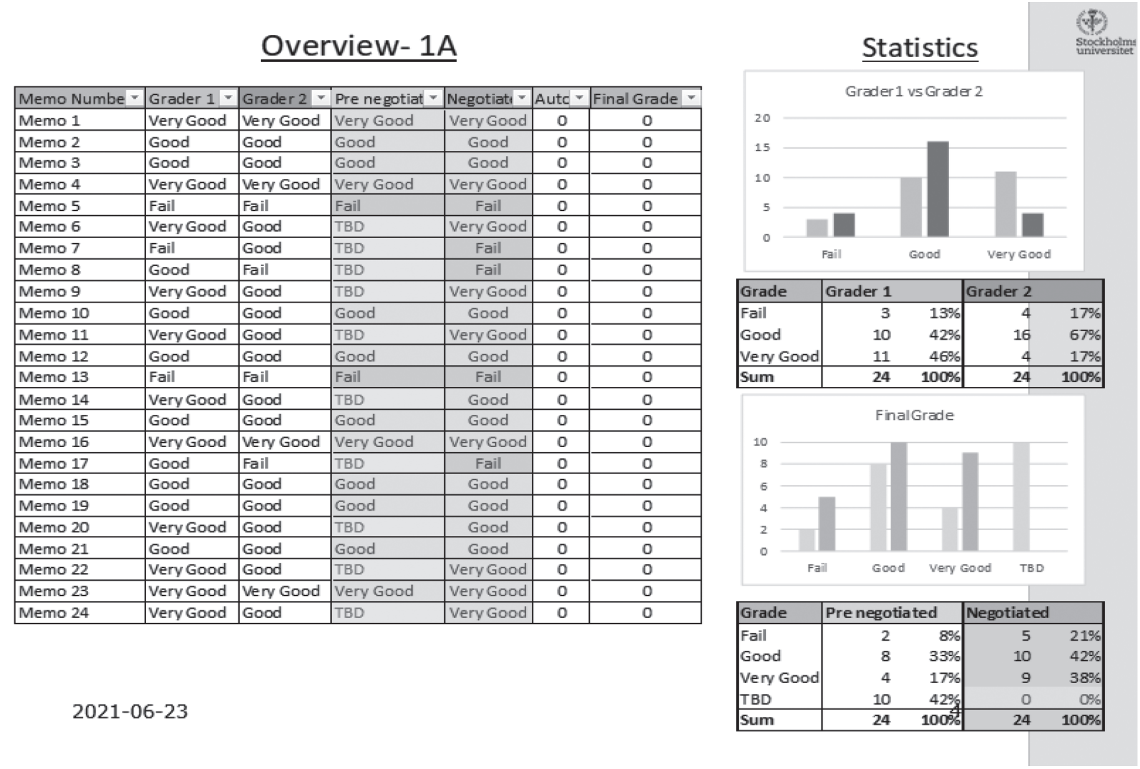
Currently on the course website 2020

Information on a research project

With the aim to free up time for implementation of the so-called Stockholm model, which focuses on the importance of critical thinking, we are within legal informatics studying which learning processes can be digitalised in a meaningful, secure and effective way. Thus far, we have introduced digital lectures and digital home exams. We are now studying the legal and technical conditions for fully or partially automated assessments of the method memos regarding the data protection regulation that are an obligatory part of the first block in the legal informatics course.

- The plan is to compare our teachers' manual assessments of the memos with what can be achieved through machine learning. The research project in question, MMM (Means for Memo Matching), is conducted in collaboration with the Department of Computer and Systems Sciences within Stockholm University. The aim is ultimately to improve the course for the students and free up time for other aspects. If we find that it is possible to use fully or partially automated assessments of the memos in a legal and suitable way, the teaching hours allotted to the course can be used to provide other meaningful education to the students (rather than being spent on assessing memos).
- Protection of the students' privacy is paramount, for which reason all texts will be pseudonymised (deidentified), to avoid the possibility of identifying any individual person.
- If you have any questions on the research project, you are welcome to contact [the course director].

Annex (B) Overview – 1A & 1B



Overview- 1A

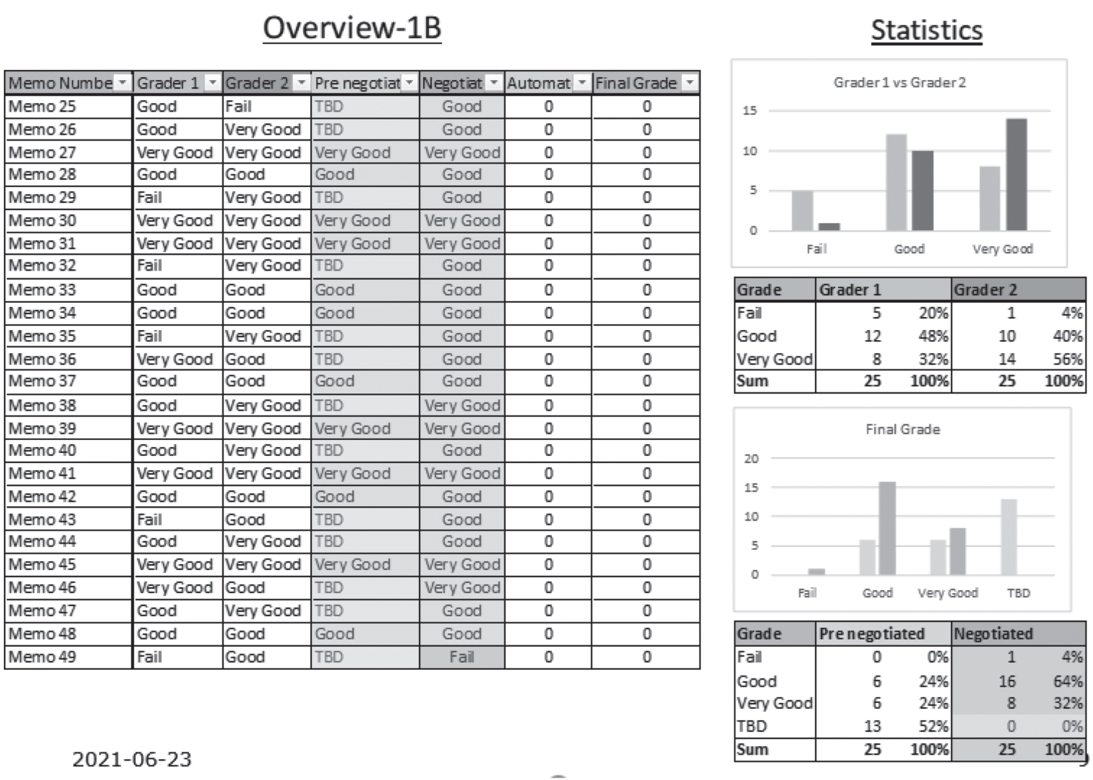
Memo Numbe	Grader 1	Grader 2	Pre negotiat	Negotiat	Autc	Final Grade
Memo 1	Very Good	Very Good	Very Good	Very Good	0	0
Memo 2	Good	Good	Good	Good	0	0
Memo 3	Good	Good	Good	Good	0	0
Memo 4	Very Good	Very Good	Very Good	Very Good	0	0
Memo 5	Fail	Fail	Fail	Fail	0	0
Memo 6	Very Good	Good	TBD	Very Good	0	0
Memo 7	Fail	Good	TBD	Fail	0	0
Memo 8	Good	Fail	TBD	Fail	0	0
Memo 9	Very Good	Good	TBD	Very Good	0	0
Memo 10	Good	Good	Good	Good	0	0
Memo 11	Very Good	Good	TBD	Very Good	0	0
Memo 12	Good	Good	Good	Good	0	0
Memo 13	Fail	Fail	Fail	Fail	0	0
Memo 14	Very Good	Good	TBD	Good	0	0
Memo 15	Good	Good	Good	Good	0	0
Memo 16	Very Good	Very Good	Very Good	Very Good	0	0
Memo 17	Good	Fail	TBD	Fail	0	0
Memo 18	Good	Good	Good	Good	0	0
Memo 19	Good	Good	Good	Good	0	0
Memo 20	Very Good	Good	TBD	Good	0	0
Memo 21	Good	Good	Good	Good	0	0
Memo 22	Very Good	Good	TBD	Very Good	0	0
Memo 23	Very Good	Very Good	Very Good	Very Good	0	0
Memo 24	Very Good	Good	TBD	Very Good	0	0

2021-06-23

Statistics

Grade	Grader 1		Grader 2	
Fail	3	13%	4	17%
Good	10	42%	16	67%
Very Good	11	46%	4	17%
Sum	**24**	**100%**	**24**	**100%**

Grade	Pre negotiated		Negotiated	
Fail	2	8%	5	21%
Good	8	33%	10	42%
Very Good	4	17%	9	38%
TBD	10	42%	0	0%
Sum	**24**	**100%**	**24**	**100%**



Overview-1B

Memo Numbe	Grader 1	Grader 2	Pre negotiat	Negotiat	Automat	Final Grade
Memo 25	Good	Fail	TBD	Good	0	0
Memo 26	Good	Very Good	TBD	Good	0	0
Memo 27	Very Good	Very Good	Very Good	Very Good	0	0
Memo 28	Good	Good	Good	Good	0	0
Memo 29	Fail	Very Good	TBD	Good	0	0
Memo 30	Very Good	Very Good	Very Good	Very Good	0	0
Memo 31	Very Good	Very Good	Very Good	Very Good	0	0
Memo 32	Fail	Very Good	TBD	Good	0	0
Memo 33	Good	Good	Good	Good	0	0
Memo 34	Good	Good	Good	Good	0	0
Memo 35	Fail	Very Good	TBD	Good	0	0
Memo 36	Very Good	Good	TBD	Good	0	0
Memo 37	Good	Good	Good	Good	0	0
Memo 38	Good	Very Good	TBD	Very Good	0	0
Memo 39	Very Good	Very Good	Very Good	Very Good	0	0
Memo 40	Good	Very Good	TBD	Good	0	0
Memo 41	Very Good	Very Good	Very Good	Very Good	0	0
Memo 42	Good	Good	Good	Good	0	0
Memo 43	Fail	Good	TBD	Good	0	0
Memo 44	Good	Very Good	TBD	Good	0	0
Memo 45	Very Good	Very Good	Very Good	Very Good	0	0
Memo 46	Very Good	Good	TBD	Very Good	0	0
Memo 47	Good	Very Good	TBD	Good	0	0
Memo 48	Good	Good	Good	Good	0	0
Memo 49	Fail	Good	TBD	Fail	0	0

2021-06-23

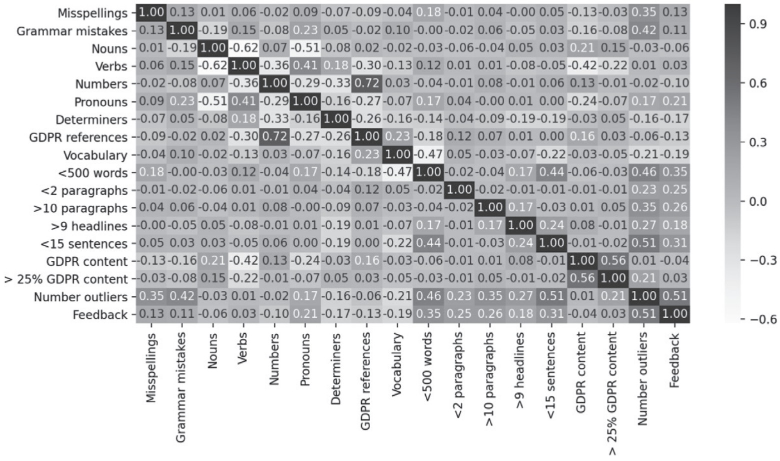
Statistics

Grade	Grader 1		Grader 2	
Fail	5	20%	1	4%
Good	12	48%	10	40%
Very Good	8	32%	14	56%
Sum	25	100%	25	100%

Grade	Pre negotiated		Negotiated	
Fail	0	0%	1	4%
Good	6	24%	16	64%
Very Good	6	24%	8	32%
TBD	13	52%	0	0%
Sum	25	100%	25	100%

Annex (C) Correlations between features and the 'fail' category

The correlation between the feedback 'fail' and each feature can be seen in the last row. *Nouns, verbs, numbers, pronouns* and *determiners* represent how common each of these parts of speech is in each student text. *General Data Protection Regulation references* correspond to the number of references made to specific articles in the General Data Protection Regulation and *Vocabulary* corresponds to the number of different words used in each text. The remaining features are 'outlier' features, except *General Data Protection Regulation content*, which shows a (weak) correlation between getting a passing grade and a high degree of *General Data Protection Regulation content*.



	Misspellings	Grammar mistakes	Nouns	Verbs	Numbers	Pronouns	Determiners	GDPR references	Vocabulary	<500 words	<2 paragraphs	>10 paragraphs	>9 headlines	<15 sentences	GDPR content	> 25% GDPR content	Number outliers	Feedback
Misspellings	1.00	0.13	0.01	0.06	-0.02	0.09	-0.07	-0.09	-0.04	0.18	-0.01	0.04	-0.00	0.05	-0.13	-0.03	0.35	0.13
Grammar mistakes	0.13	1.00	-0.19	0.15	-0.08	0.23	0.05	-0.02	0.10	-0.00	-0.02	0.06	-0.05	0.03	-0.16	-0.08	0.42	0.11
Nouns	0.01	-0.19	1.00	-0.62	0.07	-0.51	-0.08	0.02	-0.02	-0.03	-0.06	-0.04	0.05	0.03	0.21	0.15	-0.03	-0.06
Verbs	0.06	0.15	-0.62	1.00	-0.36	0.41	0.18	-0.30	-0.13	0.12	0.01	0.01	-0.08	-0.05	-0.42	-0.22	0.01	0.03
Numbers	-0.02	-0.08	0.07	-0.36	1.00	-0.29	-0.33	0.72	0.03	-0.04	-0.01	0.08	-0.01	0.06	0.13	-0.01	-0.02	-0.10
Pronouns	0.09	0.23	-0.51	0.41	-0.29	1.00	-0.16	-0.27	-0.07	0.17	0.04	-0.00	0.01	0.00	-0.24	-0.07	0.17	0.21
Determiners	-0.07	0.05	-0.08	0.18	-0.33	-0.16	1.00	-0.26	-0.16	-0.14	-0.04	-0.09	-0.19	-0.19	-0.03	0.05	-0.16	-0.17
GDPR references	-0.09	-0.02	0.02	-0.30	0.72	-0.27	-0.26	1.00	0.23	-0.18	0.12	0.07	0.01	0.00	0.16	0.03	-0.06	-0.13
Vocabulary	-0.04	0.10	-0.02	-0.13	0.03	-0.07	-0.16	0.23	1.00	-0.47	0.05	-0.03	-0.07	-0.22	-0.03	-0.05	-0.21	-0.19
<500 words	0.18	-0.00	-0.03	0.12	-0.04	0.17	-0.14	-0.18	-0.47	1.00	-0.02	-0.04	0.17	0.44	-0.06	-0.03	0.46	0.35
<2 paragraphs	-0.01	-0.02	-0.06	0.01	-0.01	0.04	-0.04	0.12	0.05	-0.02	1.00	-0.02	-0.01	-0.01	-0.01	-0.01	0.23	0.25
>10 paragraphs	0.04	0.06	-0.04	0.01	0.08	-0.00	-0.09	0.07	-0.03	-0.04	-0.02	1.00	0.17	-0.03	0.01	0.05	0.35	0.26
>9 headlines	-0.00	-0.05	0.05	-0.08	-0.01	0.01	-0.19	0.01	-0.07	0.17	-0.01	0.17	1.00	0.24	0.08	-0.01	0.27	0.18
<15 sentences	0.05	0.03	0.03	-0.05	0.06	0.00	-0.19	0.00	-0.22	0.44	-0.01	-0.03	0.24	1.00	-0.01	-0.02	0.51	0.31
GDPR content	-0.13	-0.16	0.21	-0.42	0.13	-0.24	-0.03	0.16	-0.03	-0.06	-0.01	0.01	0.08	-0.01	1.00	0.56	0.01	-0.04
> 25% GDPR content	-0.03	-0.08	0.15	-0.22	-0.01	-0.07	0.05	0.03	-0.05	-0.03	-0.01	0.05	-0.01	-0.02	0.56	1.00	0.21	0.03
Number outliers	0.35	0.42	-0.03	0.01	-0.02	0.17	-0.16	-0.06	-0.21	0.46	0.23	0.35	0.27	0.51	0.01	0.21	1.00	0.51
Feedback	0.13	0.11	-0.06	0.03	-0.10	0.21	-0.17	-0.13	-0.19	0.35	0.25	0.26	0.18	0.31	-0.04	0.03	0.51	1.00

Santa Slokenberga

EU Regulatory Responses to Medical Machine Learning in Pediatric Care: A Missed Opportunity to Overcome a Therapeutic Gap?

1 Introduction

The child's right to health has the same four core components, as anyone else's right to health – availability, accessibility, acceptability, and quality of medical care.[1] The right to health,[2] coupled with the right to enjoy the benefits of scientific progress and its applications,[3] is supposed to incentivize states to further the development of medical care and ensure that children as a group, at least in theory, do not lag behind, and that when the care is rendered available to children, it is of appropriate (and proven) quality. Reality, however, can be rather different. For years, children in general and different specific pediatric patient groups have been marginalized by being subject to "adjusted", more commonly known as

[1] See UN General Assembly, International Covenant on Economic, Social and Cultural Rights (ICESCR), 16 December 1966, United Nations, Treaty Series, vol. 993, p. 3, article 12 and UN Committee on Economic, Social and Cultural Rights (CESCR), General Comment No. 14: The Right to the Highest Attainable Standard of Health (Art. 12 of the Covenant), 11 August 2000, E/C.12/2000/4, para. 12; UN General Assembly, Convention on the Rights of the Child, 20 November 1989, United Nations, Treaty Series, vol. 1577, p. 3, article 24 and UN Committee on the Rights of the Child (CRC), General comment No. 15 (2013) on the right of the child to the enjoyment of the highest attainable standard of health (art. 24), 17 April 2013, CRC/C/GC/15, para. 112.

[2] ICESCR (n. 1) article 12.

[3] ICESCR (n. 1) article 15(1)(b).

off-label, medical care. While it is not necessarily a synonym to substandard care, quality concerns in off-label care are not foreign.[4]

Artificial intelligence (AI), including its subtype, machine learning (ML), has been labelled as "the most transformative technology" of the 21st century,[5] and is commonly hailed as a potential cure for many of the world's grave problems. AI, and in particular its subtype, medical machine learning (MML) holds the potential to transform medical care in general,[6] and pediatric medical care as its subset.[7] In addition to the tangible benefits of improved care to particular patient groups, considerable economic gain for the public health systems is forecasted.[8] Currently, considerable research initiatives are being funded to further the availability of high-quality medical care.[9] The pediatric patient population is one of the potential beneficiaries of these advances. This benefit could take place not only in terms of developing new and better diagnostic tools and cures, but also in terms of moving beyond the long-existed practices of adapting the adult patient studies and tailored outcomes for the pediatric group or subgroups therein and adjusting the therapies to match the medical needs of "smaller bodies", at least with respect to medical devices.

[4] See in that regard Santa Slokenberga, 'The Standard of Care and Implications for Paediatric Decision-Making: The Swedish Viewpoint' in Clayton Ó Néill and others (eds), *Routledge Handbook of Global Health Rights* (Routledge 2021).

[5] Daniel Schönberger, 'Artificial Intelligence in Healthcare: A Critical Analysis of the Legal and Ethical Implications' (2019) 27 International Journal of Law and Information Technology 171, 171.

[6] Although in medical care any of the mentioned techniques could be of relevance, often ML is singled out as a technique of particular importance in healthcare. Danton S Char, Nigam H Shah and David Magnus, 'Implementing Machine Learning in Health Care—Addressing Ethical Challenges' (2018) 378 The New England journal of medicine 981. Jenna Wiens and others, 'Do No Harm: A Roadmap for Responsible Machine Learning for Health Care' (2019) 25 Nature Medicine 1337. S Dolley, 'Big Data Solution to Harnessing Unstructured Data in Healthcare' [2015] IBM Report.

[7] Keeley LaForme, 'How Machine Learning Experts and Physicians at Johns Hopkins All Children's Are Working Toward a Hea' <https://www.hopkinsallchildrens.org/ACH-News/General-News/How-Machine-Learning-Experts-and-Physicians-at-Joh> accessed 6 September 2021.

[8] European Commission. Joint Research Centre. and E Gomez Gutierrez, 'Artificial Intelligence in Medicine and Healthcare: Applications, Availability and Societal Impact.' (Publications Office 2020) 13 <https://data.europa.eu/doi/10.2760/047666> accessed 6 September 2021.

[9] See section 3.

While AI applications in medicine are already a clinical reality in modern medicine, its peak generally and in pediatric medicine specifically is yet to be achieved. To enable the potential of MML to transform pediatric medical care, adequate preconditions that further scientific research in the field and lead to qualitative pediatric care need to be in place. Member States of the EU are, to a considerable degree, preempted from acting on their own in the field. The EU regulates two central elements pertaining to pediatric MML – data protection and medical devices – , and in a near future, it is expanding its regulatory grip to capture another dimension, the artificial intelligence component. Thus, even if the Member States had an interest in taking greater steps in the field, those steps would need to be subordinated to the EU preemptive actions in the field.

This contribution aims to examine the legal preconditions for developing MML tools in pediatric medicine that are set forth within the EU law. It begins by reflecting in greater detail on the AI and MML potential to transform health care generally and pediatric care as a marginalized healthcare area specifically. Thereafter, it moves on to charter key areas of EU law of relevance to the development of pediatric MML devices. Finally, it synthesizes the findings and reflects on the EU legal preconditions and the potential to overcome a gap in pediatric medical care through advancing pediatric MML. I argue that considerable steps have already been taken in order to ensure that qualitative MML medical devices for pediatric care are placed on the market, but whether these will lead to high-quality products in the field will depend on a number of considerations, including the national application of the harmonized EU requirements. I argue that none of the surveyed legal instruments, regulating data protection, medical devices and AI, contribute to furthering the development and availability of the devices directly and thus the EU misses a chance to contribute to reducing the therapeutic gap in pediatric medical care.

2 AI, MLL, medical care and pediatric medicine

AI has been broadly defined as the "science and engineering of making intelligent machines".[10] ML is a particular technique of data analytics. It can be used to develop algorithms that can learn, identify patterns, and act on the available data.[11] One central aspect of machine learning is the availability of adequate data for the algorithm to learn without being programmed to reach a particular conclusion. The learning may be either supervised or unsupervised.[12] In the former, humans label the data used to train and validate an algorithm in advance, whereas in the latter the algorithm learns patterns from unlabeled data.[13] Depending on the design, ML algorithms can be divided into two groups, the locked ones that are unable to develop themselves further and the adaptive ones that could continuously learn from data and optimize their performance.[14] The ability to learn from real-world use and experience and the capability to improved ML performance has been highlighted as one of the greatest benefits of the technology for the field of medicine.[15]

AI applications in medicine are not a novelty. The initial efforts, that date back to the 1960s, focused on diagnosis and selection of the most appropriate therapy. A prime example of such technology is a computer-based consultation technology in clinical therapeutics, the MYCIN system developed in 1972 at Stanford University. This system was developed using clinical decision criteria acquired from experts to advise

[10] 'What Is AI? / Basic Questions' <http://jmc.stanford.edu/artificial-intelligence/what-is-ai/> accessed 6 September 2021.

[11] Rohan Bhardwaj, Ankita R Nambiar and Debojyoti Dutta, 'A Study of Machine Learning in Healthcare', *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)* (2017) 237.

[12] However, technically, more nuanced approaches exist, see Batta Mahesh, 'Machine Learning Algorithms-A Review' (2020) 9 International Journal of Science and Research 381.

[13] Mahesh (n. 12).

[14] Food and Drug Administration, 'Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback'.

[15] Food and Drug Administration, 'Artificial Intelligence and Machine Learning in Software as a Medical Device' <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> accessed 4 October 2021.

physicians concerning the antimicrobial therapy section.[16] While this particular system was never used in clinical care, it was one of those contributing to furthering public interest in the field by the early 1980s, along which followed commercial interest and funding for the field.[17] However, similarly as in other fields, also in the field of medicine, AI winter came, considerably affecting interest and slowing down the field.[18] Today, new applications and startups in the field emerge on a daily basis,[19] in the aspiration to harness the benefits that the technology holds.

Building on the steady progress in the field, in the last decades, AI in medicine, and in particular MML has considerably evolved. Researchers are actively pursuing opportunities to realize the AI potential throughout all aspects of care in order to enhance and improve patient care, for example, through improved accuracy and efficiency of diagnostics, selection of therapies and prediction of outcomes.[20] There are high expectations that the field will considerably accelerate and, while caution also exists, there is considerable awareness of the importance of the technology in future medical care.[21]

Largely, AI applications in healthcare can be clustered in the following three groups: predictive use, diagnostic use and clinical decision-making aid. Predictive use of MML refers to setting a prognosis for an individual patient. Diagnostic use of MML refers to, for example, the technology assisting in arriving at the correct diagnosis for a particular patient. Clinical decision-making aids of MML refers to e.g. determining the clinical steps that need to be taken for an individual with a particular medical

[16] Edward H Shortliffe and others, 'Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System' (1975) 8 Computers and Biomedical Research 303.

[17] Edward H Shortliffe, 'Artificial Intelligence in Medicine: Weighing the Accomplishments, Hype, and Promise' (2019) 28 Yearbook of medical informatics 257.

[18] Shortliffe (n. 17).

[19] Nabile M Safdar, John D Banja and Carolyn C Meltzer, 'Ethical Considerations in Artificial Intelligence' (2020) 122 European Journal of Radiology 108768.

[20] Sudhen B Desai, Anuj Pareek and Matthew P Lungren, 'Current and emerging artificial intelligence applications for pediatric interventional radiology' (2021) Pediatr Radiol, https://doi.org/10.1007/s00247-021-05013-y.

[21] Shortliffe (n. 17). Emily Shearer, Mildred Cho and David Magnus, 'Chapter 23 - Regulatory, Social, Ethical, and Legal Issues of Artificial Intelligence in Medicine' in Lei Xing, Maryellen L Giger and James K Min (eds), *Artificial Intelligence in Medicine* (Academic Press 2021) 458.

condition, and its severity.[22] All of these applications are of importance to enhancing pediatric diagnostics, prognostics and care. In pediatric medicine, many MML applications have already been explored, including tools for prognosis, diagnosis, and therapy of pediatric critical care;[23] pediatric radiology;[24] airway management,[25] and pediatric severe sepsis prediction[26].

The development of MML technology is complex and requires accounting for a number of questions and challenges.[27] One of the central concerns is the availability of adequate training data.[28] The aim of selecting adequate data is to not only create software that functions for its given purpose in different environments, and strikes an appropriate balance between false positive and negative cases, but also to design it in a way that it is capable of dealing with potential bias[29] and that is able to address unanticipated patient contexts (known as out of sample input).[30] Admittedly, this is not a straightforward task.[31] Challenges such as these need to be tackled with due regard to the specificities of pediatric patients as a group and different subgroups therein. Generally, this patient group

[22] Fei Jiang and others 'Artificial intelligence in healthcare: past, present and future' (2017) 2 Stroke and Vascular Neurology, doi: 10.1136/svn-2017-000101.

[23] Jon B Williams, Debjit Ghosh and Randall C Wetzel, 'Applying Machine Learning to Pediatric Critical Care Data*' (2018) 19 Pediatric Critical Care Medicine 599.

[24] Michael M Moore and others, 'Machine Learning Concepts, Concerns and Opportunities for a Pediatric Radiologist' (2019) 49 Pediatric Radiology 509.

[25] Clyde Matava and others, 'Artificial Intelligence, Machine Learning and the Pediatric Airway' (2020) 30 Pediatric Anesthesia 264.

[26] Sidney Le and others, 'Pediatric Severe Sepsis Prediction Using Machine Learning' (2019) 7 Frontiers in Pediatrics 413.

[27] For an overview see Robert Challen and others, 'Artificial Intelligence, Bias and Clinical Safety' (2019) 28 BMJ Quality & Safety 231.

[28] Challen and others (n. 27).

[29] Learned biases formed on human-related data frequently resemble human-like biases towards race, sex, religion, and many other common forms of discrimination. Daniel James Fuchs, 'The Dangers of Human-like Bias in Machine-Learning Algorithms' (2018) 2 Missouri S&T's Peer to Peer 1.

[30] Kun-Hsing Yu and Isaac S Kohane, 'Framing the Challenges of Artificial Intelligence in Medicine' (2019) 28 BMJ Quality & Safety 238. This, however, in itself might not be sufficient to deal with the possible disease pattern changes. Model updating protocols are suggested as means to tackle that challenge, see Sharon E Davis and others, 'Calibration Drift in Regression and Machine Learning Models for Acute Kidney Injury' (2017) 24 Journal of the American Medical Informatics Association 1052.

[31] Challen and others (n. 27).

is diverse and complex.[32] The different patient subgroups that the pediatric group captures have different biological preconditions that must be particularly considered when developing medical care, in addition to the disease or condition-specificities and other factors that could have an impact on the medical assessment and MLL device performance in pediatric care. The considerations, such as these, shape the data selection criteria (in addition to the common bias concerns, such as ethnicity, gender) for the development of the MML technology and fragment the landscape of available data.

3 General remarks on the EU regulatory approach

The EU seeks to become a global leader in health-related AI applications.[33] Thus, it comes as no surprise that within the EU following the aspiration set out in Article 179(1) TFEU to achieve a European research area and to encouraging it to become more competitive, considerable investments are made to realize the potential that AI and MML hold for healthcare. Already in the IMI2 programme within the framework of Horizon 2020[34] considerable focus was placed on the development of new digital health solutions. This emphasis continues under the current Horizon Europe programme.[35] In parallel, considerable policy and law-making work in the field has been done in order to tackle the chal-

[32] See Kavot Zillén, Jameson Garland and Santa Slokenberga, 'The Rights of Children in Biomedicine: Challenges Posed by Scientific Advances and Uncertainties' (Council of Europe 2017).

[33] Bruno Catteneo, 'Being Smart about Our Health: How Artificial Intelligence Can Help Transform Europe's Health Sector' (*EU Science Hub - European Commission*, 3 February 2021) <https://ec.europa.eu/jrc/en/news/being-smart-about-our-health-how-artificial-intelligence-can-help-transform-europe-s-health-sector> accessed 6 September 2021.

[34] 'IMI Mission and Objectives' (*IMI Innovative Medicines Initiative*) <http://www.imi.europa.eu/about-imi/mission-objectives> accessed 4 October 2021._'History – the IMI Story so Far' (*IMI Innovative Medicines Initiative*) <http://www.imi.europa.eu/about-imi/history-imi-story-so-far> accessed 4 October 2021.

[35] AI is one of the cornerstones of the EU Innovative Health Initiative (new proposed public-private partnership (PPP) under Horizon Europe), see The Innovative Health Initiative, 'Strategic Research and Innovation Agenda' <https://www.imi.europa.eu/sites/default/files/uploads/documents/About-IMI/IHI/IHI_SRIA_DraftJune2021.pdf> accessed 4 October 2021. 'EU to Set up New European Partnerships' (*European Commis-*

lenges and leash the potential that AI generally[36] and AI in the health field holds.[37]

Realizing this EU aspiration for a competitive position in health-related AI and simultaneously furthering the Member States' responsibility regarding children's health rights, that is, the right to health and the right to benefit of scientific advances, as formulated by the UN in the ICESCR, relates to several areas where the EU has legislative competencies. Neither of them is sufficient in themselves for realizing the potential that MML holds for pediatric medicine (i.e. lack of exclusive competence), and complementary actions on the part of the Member States might be needed. However, taken together, they capture key elements and consequently preempt the Member States from the action in the field. Thus, a careful navigation on the principle of conferral and the domain of shared and complementary competence and an investigation of how it has been exercised is necessary.

To begin with, the development of pediatric AI-based medical devices requires the availability of a sufficient amount of appropriate health data. As previously noted, the pediatric patient group is fragmented, and consequently so is the pediatric data landscape. The current central framework in that regard is the General Data Protection Regulation (GDPR).[38] It sets out rules to safeguard the rights of natural persons, whilst also ensuring free movement of personal data. In regard to scientific research, considerable discussions have emerged regarding the GDPR's potential negative impact, e.g. data sharing hurdles that are now attributable to the

sion – European Commission) <https://ec.europa.eu/commission/presscorner/detail/en/ip_21_702> accessed 4 October 2021.

[36] See, for example, 'Expert Group on AI | Shaping Europe's Digital Future' <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai> accessed 7 October 2021.

[37] Recently, also the ethical and societal impacts of the AI in medicine and healthcare technology have been chartered to inform policy-making AI Watch Artificial Intelligence in Medicine and Healthcare: applications, availability and societal impact p. 13 European Commission. Joint Research Centre., *Artificial Intelligence in Medicine and Healthcare: Applications, Availability and Societal Impact.* (Publications Office 2020) 13 <https://data.europa.eu/doi/10.2760/047666> accessed 4 October 2021.

[38] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) OJ L 119, 4.5.2016, p. 1–88.

GDPR.[39] In response to the challenges in the field and in order to realize the potential that big data hold in the field of health and medical care, the European Health Data Space is under the development. The European Commission has committed to present its proposal in the last quarter of 2021,[40] thus at the moment of writing, it is uncertain whether and to what extent it could heal the existing challenges. Additionally, within the EU, stand-alone software for healthcare purposes could fall within the scope of a definition of a medical device provided in Article 2(1) of the Medical Devices Regulation (MDR).[41] Consequently, it shall comply with the detailed quality requirements prescribed by the regulation. Finally, the European Commission has put forward a proposal for regulating AI (proposed AIR).[42] Thus, in a near future, once the proposed AIR passes the legislative stage and enters into force also those requirements will need to be followed.

4 GDPR and development of pediatric AI tools

The development of AI/MLL medical devices intended for pediatric care is a data-intensive activity. There are a number of data sources that can cater to the need. For example, data may be obtained from research repositories, such as biobanks, or from the clinical or public health environment, for example, electronic health records or public insurance data. Data that may also be obtained non-clinically, for example, collected

[39] See, for example, Santa Slokenberga, ‘Setting the Foundations: Individual Rights, Public Interest, Scientific Research and Biobanking’ in Santa Slokenberga, Olga Tzortzatou and Jane Reichel (eds), *GDPR and Biobanking: Individual Rights, Public Interest and Research Regulation across Europe* (Springer International Publishing 2021).

[40] Vincent Draguet, ‘European Health Data Space’ (*Public Health - European Commission*, 18 September 2020) <https://ec.europa.eu/health/ehealth/dataspace_en> accessed 4 October 2021.

[41] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance) OJ L 117, 5.5.2017, p. 1–175. See Article 2(1) for the definition.

[42] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 final.

from wearables or mobile devices and be used in the development of MML devices.[43]

The General Data Protection Regulation (GDPR) could be said to have a rather complex relationship with the development of pediatric MML tools. Generally, the GDPR per se is not a research regulatory instrument.[44] However, it contains a research regime and prescribes particular rules that shall be followed when personal data are processed. The development of MML devices for pediatric care in the EU commonly will risk triggering the application of the GDPR, except for when anonymized data are used for the purpose.[45] While there could be parts of a study that could be performed, using anonymized data and thus it would not trigger the application of the GDPR, there will be MML research that is difficult to perform anonymously, in either full or in part. Moreover, even if anonymity might seem an option, the degree of anonymity could be an illusion.[46] In cases such as those, the application of the GDPR is rather unavoidable.[47]

The GDPR does not further any particular type of scientific research in a direct way. However, at the same time, the development of pediatric MML devices falls under the broader spectrum of public interest in the area of public health.[48] The domain of public interest, both separately and within the area of scientific research, can be subject to further particularly lax requirements.[49] Thus, although the development of pediatric MML devices is located within the general research regime, it has potential to benefit from the specific public interest rules.

In order for the development of a pediatric AI tool to take place, the particular data processing activities underlying its development needs to be regarded as lawful within the meaning of the GDPR. This entails meeting the general requirement for a lawful basis for personal data processing under Article 6(1) as well as the specific requirement in relation to sensitive personal data, which includes health data, under Article 9(2). Article 6(1) does not treat scientific research in any special way and thus,

[43] Shearer, Cho and Magnus (n. 21) 459.

[44] Slokenberga, 'Setting the Foundations' (n. 39) 17.

[45] See Recital 26 and Article 2(1) GDPR.

[46] Paul Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' (2009) 57 Ucla L. Rev. 1701.

[47] See Article 2(1) GDPR.

[48] Recital 159 GDPR.

[49] See Slokenberga, 'Setting the Foundations' (n. 39) 11–27.

the research activity needs to meet one of the general legal grounds, such as consent or public interest, set forth in that provision. Article 9(2), however, does treat scientific research specifically in Article 9(2)j, where processing that is necessary for scientific research purposes is explicitly mentioned as one of the exceptions to the general prohibition (Article 9(1)) to process sensitive data This does not mean that other legal grounds are irrelevant, especially, if the freedom that is afforded to the Member States under Article 9(4) GDPR is accounted for. This provision *expressis verbis* allows the Member States to maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health. These further regulations could theoretically be such that they treat pediatric AI research in a special way if this treatment is shaped in line with the general principles of EU law. However, a study carried out by the EU does not show that thus far it has been used for that purpose.[50]

While the GDPR prescribes several alternative legal grounds that could be used for the processing of personal data in developing MML devices for pediatric patients, not all of them might be equally suitable or desirable for the purpose. Generally, in the field of scientific research, and in particular, in the aftermath of the WWII, the principle of informed consent has been of paramount importance. While, as explicitly illustrated by the GDPR rules, data-driven research has drifted from that approach, informed consent is still commonly recognized as a safeguard in processing personal data for research purposes, disregarding whether it is used as legal basis for data processing.[51] The GDPR enables several alternative legal basis for the processing of personal data, however, preference of another legal basis over consent could be regarded as sensitive and in conflict with the sense of ownership that people could have about their data, even if this sense of ownership is rather illusory.[52] The tensions that using other legal basis instead of consent is in conflict with the sense of ownership people could have regarding their data (or responsibility for

[50] DG Health and Food Safety, 'Assessment of the EU Member States' Rules on Health Data in the Light of GDPR' (European Union 2021).

[51] GDPR (n. 38) recital 33. See also Ciara Staunton, Santa Slokenberga and Deborah Mascalzoni, 'The GDPR and the Research Exemption: Considerations on the Necessary Safeguards for Research Biobanks' (2019) 27 European Journal of Human Genetics 1159.

[52] Santa Slokenberga, 'You Can't Put the Genie Back in the Bottle : On the Legal and Conceptual Understanding of Genetic Privacy in the Era of Personal Data Protection in Europe' [2021] Biolaw Journal - Rivista di Biodiritto 223.

their child's data), and could be illustrated by such case as the Italy – IBM Watson Health case and Google Health – NHS case, even though both occurred pre-GDPR application period and one can question how these cases would evolve had they occurred under the GDPR. Regarding the former, it has been reported that detailed medical records of 61 million Italian citizens have been provided to IBM Watson Health by the Italian government.[53] Among the concerns that have been highlighted is lack of consent of the persons concerned.[54] While the data scale was different in the Google Health – NHS case, also this transfer brought along considerable questions regarding the collaboration, and lack of the data subjects' control over the transfer through their consent.[55]

Further to the legal basis through which the development of MML for pediatric care could be furthered, also other research-furthering legal interventions are possible under the GDPR. One such is the derogation from individual rights that the GDPR provides for the data subjects in Chapter III. The extent of derogations will depend on a particular activity and circumstances surrounding the research. However, more generally it could be said that the GDPR enables a two-level derogation avenue.[56] First, through the direct application of the individual rights provisions that permit derogations through Article 89(1) GDPR, and secondly, through national law or EU law that prescribes particular derogations in accordance with Article 89(2) GDPR. While Article 89(2) in itself does not explicitly enable the Member States or the EU to treat a particular research field specifically, it could be done considering the discretion that the Member States have under Article 9(4) GDPR, and in so far as it is exercised in line with the fundamental principles of EU law. More-

[53] 'Detailed Medical Records of 61 Million Italian Citizens to Be given to IBM for Its "Cognitive Computing" System Watson' (*PIA VPN Blog*, 22 May 2017) <https://www.privateinternetaccess.com/blog/detailed-medical-records-61-million-italian-citizens-given-ibm-cognitive-computing-system-watson/> accessed 4 October 2021.

[54] Janos Meszaros, Marcelo Corrales Compagnucci and Timo Minssen, 'The Interaction of the Medical Device Regulation and the GDPR: Do European Rules on Privacy and Scientific Research Impair the Safety & Performance of AI Medical Devices?' in Glenn I Cohen and others (eds), *The Future of Medical Device Regulation: Innovation and Protection* (Cambridge University Press (available at Social Science Research Network) 2021) <https://papers.ssrn.com/abstract=3808384> accessed 4 October 2021.

[55] Julia Powles and Hal Hodson, 'Google DeepMind and Healthcare in an Age of Algorithms' (2017) 7 Health and Technology 351.

[56] Considering Article 23 GDPR derogations, a three-level avenue. Slokenberga, 'Setting the Foundations' (n. 39).

over, this freedom would also need to be considered accounting for other circumstances, for example, Member State's international commitments and research regulatory traditions.

MML for pediatric care and public interest concept adds further nuance and possibilities to further scientific research and development of technology for this patient group.[57] To begin with, already recital 159 GDPR indicates that research that is carried out in the field of public health is in the public interest. As such, it could benefit from a particular legal basis under Article 9(2) GDPR as well as other specific derogations that are prescribed in the GDPR. At the extreme, if the degree of public interest reaches the level of an "important objective of general public interest", Article 23(1)(e) GDPR could be triggered, which enables far-reaching derogations from the GDPR rules. There are, however, several challenges with the application of the public interest rule under the GDPR. For example, there is a lack of detailed guidance on the concept of public interest as well as on the threshold to measure and classify the level of interest.[58] It is thus uncertain not only what level of public interest MML in pediatric medicine could trigger, but also whether all MML in the field should be treated the same way, or pediatric MML, given the challenges in pediatric medicine generally, could be subjected to particular rules.[59]

5 AI in pediatric care and development of medical devices

The Medical Devices Regulation (MDR) subjects any software intended for the prediction or prognosis of a disease or monitoring of treatment,[60] to its requirements. Substantively, the regulation sets forth rules concerning the placing on the market, making available on the market or putting into service of MML pediatric devices.[61] The EU does not claim any general competence over the regulation of the clinical investigation of med-

[57] However, as elsewhere has been noted, what public interest is not something that is well-elaborated within the GDPR, see Slokenberga, 'Setting the Foundations' (n. 39).

[58] Recital 73 affirms the wording of Article 23(1)(e) GDPR.

[59] Meszaros, Corrales Compagnucci and Minssen (n. 54) 3.

[60] MDR (n. 41) article (1).

[61] MDR (n. 41) article 1(1).

ical devices.[62] However, where a clinical investigation is carried out as part of the clinical evaluation for conformity assessment purposes it shall meet the requirements prescribed by the regulation.[63] This could be said to be a rather technical regulatory approach to hide a larger pre-empting policy behind the EU's teeth, as any software placed on the market, made available on the market or put into service within the EU shall meet the requirements set out in the regulation.[64] Thus, a medical device that does not meet the prescribed requirements cannot be made available to the users.[65]

As a starting point, a pediatric MML device shall meet the applicable general safety and performance requirements.[66] Identification of the relevant general safety and performance requirements that a particular MML pediatric device shall meet under the MDR lies with the manufacturer. Moreover, manufacturers shall also specify and justify the level of clinical evidence that is necessary in order to verify whether they are met.[67] Generally, any medical device shall be suitable for the intended purpose. Moreover, the device shall be safe and effective and shall not compromise the clinical condition or the safety of patients, healthcare personnel or other persons. While the general requirements aspire to ensure a high level of protection of health and safety, they also acknowledge that some risks could be deemed acceptable when weighed against the benefits to the patient.[68]

In addition to the general safety and performance requirements, other requirements also apply, given the nature of a particular medical device. While the regulation does not prescribe any particular requirements re-

[62] It is defined as "any systematic investigation involving one or more human subjects, undertaken to assess the safety or performance of a device", MDR (n. 41) article 2(45).

[63] MDR (n. 41) article 62(1).

[64] The application of the clinical investigation requirements is subjected to meeting at least one of the further in the regulation specified applicability requirements in article 62(1) of the Regulation. Those are as follows: either 1) to establish and verify that the device in question achieves the performance intended as specified by its manufacturer, or 2) to establish and verify the clinical benefits of a device as specified by its manufacturer, or 3) to establish and verify the clinical safety of the device and to determine any undesirable side-effects, and assess whether they constitute acceptable risks when weighed against the benefits to be achieved by the device.

[65] See MDR (n. 41) article 5.

[66] MDR (n. 41) article 5(2).

[67] MDR (n. 41) article 61(1). For the requirements see annex I.

[68] See MDR (n. 41) annex I.

garding medical devices containing a ML component specifically or AI component generally, it sets relevant requirements that are applicable to any software. Thus, for example, software, such as pediatric MML, shall be designed to ensure repeatability, reliability and performance in line with their intended use. Moreover, it shall be developed and manufactured in accordance with the state of the art considering the principles of the development life cycle, risk management, including information security, verification and validation. The manufacturer shall also determine minimum requirements concerning hardware, IT networks characteristics and IT security measures. Finally, when software is intended to be used in combination with mobile computing platforms, compatibility requirements e.g. size and a contrast ration of a screen, as well as different environments need to be accounted for.[69]

The verification procedure of the requirements that MML pediatric devices will need to meet under the MDR depends on the classification of a particular device. The higher the risks associated with a particular device vis-à-vis the human body,[70] the higher the classification it has. The regulation does not prescribe any particular requirements for the MML devices intended for pediatric patients. The general requirements apply. As derives from Rule 11 of Annex VIII of the Regulation, any software that does not fall within a particular exception specified within this rule shall be classified as a class I device.[71] Exceptions are as follows. Software that is intended to provide information that is used to make decisions with diagnosis or therapeutic purposes is classified as class IIa device. However, if such decisions have an impact that may cause death or an irreversible deterioration of a person's state of health, then the device shall be classified as a class III device. In case such decisions have an impact that may cause serious deterioration of a person's state of health or surgical intervention, then the device shall be classified as a class IIb device. Generally, software that is intended to monitor physiological processes is classified as class IIa; however, if it is intended for monitoring of vital physiological parameters, where the nature of variations of those parameters is such that it could result in immediate danger to the patient, it is classified as class IIb.

[69] MDR (n. 41) annex I, chapter II, section 17.

[70] See MDR (n. 41) recital 58. See also annex VIII, chapter II, section 3.1.

[71] See MDR (n. 41) article 51 and annex VII, rule 11.

The classification particularly relates to different procedures for the conformity assessment of the devices, and the degree of involvement of the notified body in the assessment.[72] Class I devices are associated with a low level of vulnerability. Hence, the conformity assessment procedure should be carried out under the sole responsibility of manufacturers.[73] For class IIa, IIb and class III devices, a particular level of involvement of a notified body is mandated.[74] The involvement of a notified body seeks to ensure that it has assessed the conformity to the applicable requirements under the MDR; however, the modalities of the involvement differ for different device classes and the MDR leaves some room for choice regarding the exact procedure that is to be followed.[75] One of the key implications of a particular MML pediatric device falling beyond class I is that it will be subject to a conformity assessment carried out by a notified body. A successful assessment procedure results in a conformity assessment certificate,[76] attachment of a CE mark indicating that the applicable requirements of the MDR are met.[77]

The degree of involvement of a notified body in the conformity assessment procedure for class IIa, IIb and III devices depends on the modalities of the conformity assessment and particular audit requirements. One central aspect in all assessments that are based on a quality management system and assessment of technical documentation is audit.[78] For MML pediatric medical devices, this could include reviewing of input data that are laying at the core of the devices. This raises such questions as the existence of appropriate expertise of the notified bodies for carrying out such a review.[79] In addition to the necessary knowledge regarding a particular medical application and condition in question, also particular understanding of the pediatric patient group is necessary, along with an understanding of the risks of bias in the respective group. Moreover, the standard will be regarded as adequate in order to be considered as meet-

[72] Choice of a notified body is addressed in article 42.

[73] MDR (n. 41) recital 60.

[74] MDR (n. 41) recital 60.

[75] See MDR (n. 41) article 52 that sets out general rules on conformity assessment procedure.

[76] MDR (n. 41) article 56.

[77] MDR (n. 41) article 20.

[78] See MDR (n. 41) annex IX.

[79] Meszaros, Corrales Compagnucci and Minssen (n. 54) 3.

ing the applicable requirements is also uncertain, and how it will differ between different purposes in pediatric medicine.

6 How, if at all, the proposed AI Regulation change the landscape?

Different MML approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning, fall within the scope of the application of the proposed AI Regulation (AIR) as AI-systems.[80] Thus, any pediatric MML device is also expected to trigger the application of the proposed AIR once it is passed and enters into effect.

The proposed AIR neither enables nor facilitates a particular field of AI application. It is a standard setter; and as such, it sets uniform principles in the attempt to ensure safe and reliable AI systems within the EU.[81] In that regard, it creates a three-level risk regime. The AI systems that create an unacceptable level of risk are prohibited. The AI systems that pose a high-risk fall within a particular high-risk regime set forth within the proposed AIR. The devices that fall within the low or minimal risk regime fall under the general requirements prescribed within the proposed AIR. The high-risk AI system regime captures those AI-systems that are of importance for the functioning of the society, as well as individuals, such as medical devices.[82] This means that they are subjected to strengthened legal requirements, including regarding accuracy, robustness and cybersecurity,[83] and the requirement if a human oversight.[84]

As a general rule, any high risk devices, including pediatric MML devices may only be placed on the market or put into service if the requirements

[80] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 final, annex III.

[81] See proposed AIR (n. 80) recital 18 and article 1.

[82] See proposed AIR (n. 80) article 6 in conjunction with annex II.

[83] Proposed AIR (n. 80) article 15.

[84] Proposed AIR (n. 80) article 14.

that the proposed AIR prescribes are met.[85] As previously noted, adequate data is an essential precondition in developing MML pediatric devices. The proposed AIR sets forth requirements for the training, validation and testing data sets that are used in the development of pediatric MML medical devices as high-risk AI systems. It requires that these data sets be subject to "appropriate data governance and management practices".[86] The appropriateness is intended to include availability, quantity, as well as suitability of the data sets for the intended products.[87] The proposed AIR sets forth further requirements for the data used in developing pediatric MML medical devices. This means that the data sets that are used in the development of a device shall be relevant, representative, free of errors and complete, including with due regard to the specificities of a particular user group.[88] Moreover, these data sets shall also account for the characteristics or elements that are particular to the specific setting within which the pediatric medical device is intended to be used.[89] Thus, although the proposed AIR sets forth a framework that aim to ensure that adequate data sets are used for developing pediatric MML medical devices, what adequate is will need to be assessed based on the state of the art, as well as a particular case at hand. As the responsibility lies with the manufacturer, divergences in the approaches for managing data can be foreseen that can reflect in the technicalities and the performance of the device.

Similarly as for medical devices generally, also for AI-systems a technical documentation needs to be drawn up before a particular system is placed on the market or put into service.[90] For medical devices specif-

[85] Proposed AIR (n. 80) article 19.1. Central requirements: an iterative risk management system needs to be established, implemented and documented (article 9). In the development, datasets that meet prescribed governance and management requirements shall be used (article 10). The high-risk AI system shall have automatic recording capacity while the system is operating (article 12). It shall be designed and developed in such a way to ensure that their operation is sufficiently transparent for the users (article 13) and that they can be effectively overseen by natural persons when in use (article 14). Moreover, they shall achieve an appropriate level of accuracy, robustness and cybersecurity, and maintain this throughout their lifecycle (article 15) Finally, adequate technical documentation needs to be drafted that allows assessing compliance with the applicable requirements set out in the proposed AIR (article 11).

[86] Proposed AIR (n. 80) article 10(6).

[87] See Proposed AIR (n. 80) article 10; in particular, 10(2)(e).

[88] Proposed AIR (n. 80) article 10(3).

[89] Proposed AIR (n. 80) article 10(4).

[90] Proposed AIR (n. 80) article 18(1).

ically, one single document needs to be drawn that contains both, the information mandated under the MDR, as well as information necessary under the proposed AIR.[91] An overarching requirement is set forth, that this documentation needs to be "kept up to date" and is drafted in a way that enables ascertaining compliance with the applicable requirements set out in the proposed AIR.[92] While the minimum requirements are outlined in the proposed AIR, further requirements could apply due to the nature of the device in question. Moreover, in light of the technical progress, the European Commission has been delegated the power to specify other requirements through relevant amendments in the proposed AIR annex IV.[93]

An essential element prior to placing on the market or putting into service high risk pediatric devices is conformity assessment, an obligation that generally lies with the providers (developers) of the devices.[94] However, exceptionally, for health purposes derogation of this requirement could be possible.[95] This assessment is based on the quality management system assessment of technical documentation, including granting of the availability for the examination of the datasets used for developing the device.[96] In examining the technical documentation, the notified body is enabled to require that the provider supplies further evidence or carries out further tests to enable a proper assessment of conformity of the AI system with the applicable requirements. However, if the notified body is not satisfied with the tests carried out by the provider, the notified body shall directly carry out adequate tests, as appropriate.[97] This mean that at this stage, the notified body is entrusted with a far-reaching obligation in carrying out the assessment.

For the users, of central importance is Article 13 that sets forth transparency and provision of information requirements. It is required that AI systems are designed and developed in a way that allows ensuring that their operation is sufficiently transparent to enable the users of a device to interpret the system's output and use it appropriately.[98] Moreover, this

[91] Proposed AIR (n. 80) article 11(2).

[92] Proposed AIR (n. 80) article 11(1).

[93] Proposed AIR (n. 80) article 11(3). See article 3(1 *sic*) for the definition of a provider.

[94] See Proposed AIR (n. 80) article 19(1) and 43.

[95] Proposed AIR (n. 80) article 47.

[96] Proposed AIR (n. 80) annex VII, section 4.3.

[97] Proposed AIR (n. 80).

[98] Proposed AIR (n. 80) article 13(1).

information shall meet certain quality requirements, including regarding conciseness, completes, correctness and clarity, and be relevant, accessible and comprehensible to the intended users of the device.[99] The regulation sets forth further modalities of the information,[100] including particular requirements relevant for enabling the user to understand the performance of the device. One of the challenges in pediatric care will relate to ensuring that appropriate information is provided to the patients as users of the devices, given the peculiarities of different patient groups.

7 Concluding reflections

As becomes rather apparent, none of the three legal instruments is concerned with directly furthering the development of the pediatric MML devices. What concerns the GDPR, it can be noted that a number of mechanisms are inbuilt within the framework that enable and could even further data processing. However, a closer look at the legal basis system reveals a different image. While the GDPR sets forth various legal basis for personal data processing, it does not address preconditions for accessing these data. The Member States retain freedom to decide who and under what circumstances are able to access data for research purposes. Consequently, in itself, it is insufficient for furthering scientific research. For example, the GDPR in itself cannot be used as basis to claim access to the patient data from the national electronic health records for scientific research purposes in order that the development of pediatric MML can take place. This could, however, happen if there is, for example, appropriate national law in place securing the access.[101] This could be said to be one of the central shortcomings in developing pediatric MML, as often-multinational collaboration will be necessary to cater for the data need for a particular patient group. More broadly, it could be argued to be also one of shortcomings of the GDPR research regime generally, and something that should be carefully considered as the European Health Data Space is being developed.

[99] Proposed AIR (n. 80) article 13(2).

[100] Proposed AIR (n. 80) article 13(3).

[101] This anchors in an argument that access rights and the control over personal data shall not be reduced to the legal basis for the processing of personal data, and fulfilling respective processing preconditions.

The MDR provides a rather principle-based framework to further a high level of health and safety requirements for the MML pediatric devices. For the devices that require the involvement of the notified body, it designs a balanced mechanism whereby the manufacturer decides on the applicable requirements and ensures that they are met, whereas the notified body verifies whether the requirements are met. Whether this approach is adequate for ensuring qualitative pediatric MML devices remains to be seen. As usual, the devil is in the details, and in particular the assessment of how the applicable principles set out in the MDR play out in an individual MML pediatric device case.

The proposed AIR, the newcomer in the regulatory spectrum, generates rather complementary effects with the MDR. In particular, in regard to the applicable requirements for the performance of software as a medical device. It should then be so that data quality requirements set out in the proposed AIR are relevant for meeting the standard that the MDR requires. The synergy between two instruments is through, for example, the single technical documentation requirement, which needs to be drafted in a way to meet the requirements of both of the applicable instruments, as well as the conformity assessment mechanisms.[102] Ultimately, at least on the surface, the two instruments should generate a mechanism for ensuring qualitative pediatric MML devices in the internal market.

While the EU has preempted Member States form action in the area of pediatric MML devices to a considerable degree, in particular, regarding conditions for data processing, as well as quality requirements for medical devices and AI-systems, these mechanisms fail to generate concrete initiatives for development of these devices, and thus enhancing availability of medical care. Only the GDPR contains a mechanism that could further the development, albeit in a rather indirect way. This is in contrast with the long-existing mechanism to further the development of pediatric medical products through an internal market regulation,[103] and could risk leading to a situation where a particularly marginalized patient group is also down-prioritized in the development of new pediatric MML devices.

[102] Proposed AIR (n. 80) article 43(3).

[103] Regulation (EC) No 1901/2006 of the European Parliament and of the Council of 12 December 2006 on medicinal products for paediatric use and amending Regulation (EEC) No 1768/92, Directive 2001/20/EC, Directive 2001/83/EC and Regulation (EC) No 726/2004 (Text with EEA relevance) OJ L 378, 27.12.2006, p. 1–19.

Charlotte Högberg & Stefan Larsson

AI and Patients' Rights: Transparency and Information Flows as Situated Principles in Public Health Care

Abstract

The development of artificial intelligence (AI) for medicine and health care is rapidly evolving. However, the automation, scale and data dependency of AI-driven decision-making and decision-support calls for a reassessment of principal ethical and legal norms of transparency, in the light of these novel methodologies. The quality of AI-driven health care, we argue, is depending on it. In this chapter, we provide an overview of novelties that AI in health care bring about, in order to identify key aspects potentially affecting current legal and normative (medical ethical) principles related to transparency and explainability. We develop a conceptual framework on transparency in general and explainability in particular, in relation to AI in health care. Further, we analyse principal and normative legal frameworks of patients' rights relating to transparency and explainability – e.g., right to information, autonomy and privacy – within Sweden and the EU. Doing so, we outline main challenges in the implementation of AI in, primarily public, health care. We argue that there is an interdependency between health care quality and transparency. As transparency is not a binary state, but something that is *situated* in information practices, it is important to consider what kind of transparency is needed to safeguard the best possible health care. We find that meaningful and contextual transparency and explainability of AI-systems and methodologies is necessary to adhere to the basic principles of normative and legal frameworks of Swedish health care, including

patient autonomy. In addition, meaningful and contextual transparency is also a prerequisite for assessing if the best possible care is given to the one most in need.

1 Introduction

According to the modernized version of the Hippocratic oath, The Declaration of Geneva, a physician's main priority should be the health and well-being of the patient.[1] This ideal and other moral values – such as to respect human life, the integrity and autonomy of the patient, the patient's right to information, and to conduct care in an ethical manner and use medical knowledge for good – are considered principles for medicine and health care. These can be found in a wide array of policies, legal frameworks and guidelines. Technological innovations in drug discovery, treatments, diagnostic tools and so forth, have contributed to the fulfilment of these values and to improved chances for health and longevity. Still, the implementation of new technologies can pose ethical and legal challenges to ideals of medicine and healthcare. Technology also tends to develop faster than regulations adapt, causing *the pacing problem* – as pointed out in socio-legal studies.[2] Many of the latest technological innovations in medicine and health care are based in Artificial Intelligence (AI) and machine learning in particular. AI consists of a broad collection of technologies and methods. As described by Dignum:

> [AI] deals not only with how to represent and use complex and incomplete information logically but also with questions of how to see (vision), move (robotics), communicate (natural language, speech) and learn (memory, reasoning, classification).[3]

While AI is not new, it is now a fast-growing field due to increased access to data, computer power, and the creation of new and improved machine learning models. This is true also for AI within medicine and health care.

[1] 'Declaration of Geneva' (World Medical Association 2021) <https://www.wma.net/policies-post/wma-declaration-of-geneva/> accessed 2021-05-20.

[2] E.g., Stefan Larsson, 'AI in the EU: Ethical Guidelines as a Governance Tool', in Antonia Bakardjieva Engelbrekt, Karin Leijon, Anna Michalski & Lars Oxelheim (eds.) *The European Union and the Technology Shift*. (Palgrave Macmillan 2021).

[3] Virginia Dignum, 'Introduction' in Virginia Dignum (ed.), *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (Springer International Publishing 2019) 3.

The purpose of this chapter is to outline main transparency challenges of implementing AI in the public health sector. In order to do this, we adress the following research question:

- What is the role of transparency and explainability of artificial intelligence in relation to patients' rights and information flows in Swedish health care?

Our main methodology is a qualitative analysis of patients' rights in health care information flows, which could be affected by the use of applications based on AI. In the remainder of this section, we discuss the role of AI in healthcare. In section 2 we present a theoretical discussion of transparency and explainability in relation to AI. This theoretical discussion forms the foundation for a socio-legal analysis (section 3) of documents representing medical ethics, as exemplified by the Declaration of Geneva,[4] and an array of relevant legal frameworks at both EU level as well as at the Swedish level. The most central regulations at the EU level are the General Data Protection Regulation, GDPR,[5] in force since May 2018, and the Medical Device Regulation, MDR,[6] that is fully applicable since May 2021. In Sweden, at the national level, the main regulatory instruments are the Health and Medical Service Act,[7] the Patient Act,[8] the Patient Safety Act[9] and the Patient Data Act.[10]

Clearly, however, this list is not exhaustive. The rights of patients and the obligations of health care providers and health professionals are regulated by a large number of national and international laws, including Swedish constitutional law, as well as by extra-legal norms, ideals, global policies, common standards and local guidelines. In different ways, they concern the legitimacy of information flows. Analysing all of these and

[4] 'Declaration of Geneva' (n. 1).

[5] Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Hereinafter cited as GDPR.

[6] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. Hereinafter cited as MDR.

[7] Hälso- och sjukvårdslag (SFS 2017:30).

[8] Patientlag (SFS 2014:821).

[9] Patientsäkerhetslag (SFS 2010:659).

[10] Patientdatalag (SFS 2008:355).

their hierarchical validities and mutual relations goes beyond the scope of this chapter. Instead, the goal of this chapter is to pinpoint main informational principles within the legal and normative frameworks of medicine and health care, which may be affected by the implementation of AI. These informational principles include the right to (equal) health care (section 3.1), the right to privacy and integrity (section 3.2), the right to information (section 3.3), and the right to dignity and autonomy (section 3.4). In section 4 we discuss how an AI application with high predictive accuracy but with low transparency does not lead to the best possible care, if the lack of transparency means that the aforementioned informational principles are not adhered to. In section 5 we summarize the main argument and findings presented in this chapter.

1.1 Background: The promises and challenges of AI in health care

Medicine and health care are important fields of application for AI. The use of AI-systems and methodologies in healthcare could have a beneficial, or even vital, impact if it would result in improved predictions, diagnoses and prognoses of diseases. The broader effects could be better health, improved well-being and an increased amount of successful outcomes of treatments. A desired goal of AI implementation is also increased efficiency, especially as the health sector is also facing increased costs and administrative burdens, scarcity of practitioners and aging populations.[11] Another hope is the personalisation of medicine, as stated by experts in the field:

> Machine learning will become an indispensable tool for clinicians seeking to truly understand their patients. As patients' conditions and medical technologies become more complex, the role of machine learning will grow, and clinical medicine will be challenged to grow with it.[12]

[11] E.g., *Ethics and governance of artificial intelligence for health: WHO guidance* (World Health Organization 2021), Arash Shaban-Nejad, Martin Michalowski and David L. Buckeridge, 'Explainability and Interpretability: Keys to Deep Medicine' in Arash Shaban-Nejad, Martin Michalowski and David L. Buckeridge (eds.), *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability* (Springer International Publishing 2021).

[12] Ziad Obermeyer and Ezekiel J. Emanuel, 'Predicting the Future – Big Data, Machine Learning, and Clinical Medicine' (2016) 375 N Engl J Med 1216 1218.

One of the areas in the forefront is image analysis for radiology,[13] where AI-systems have, for example, been found to have an accuracy in cancer detection comparable to average breast radiologists.[14] AI is also used for other types of clinical decision support-tools, as well as in the form of conversational AI, offering more and faster ways of communication. It is furthermore used for administrative purposes, such as scheduling staff, allocating resources and making cost predictions.[15] Contributing to this development is the increase of digital health data and the datafication of health care.[16] The Scandinavian countries have a possible advantage due to large amounts of public health data, such as in national registers. The instated *Vision for e-health* declares that by the year 2025, Sweden should be world leading in using the opportunities provided by digitalization and e-health.[17] One factor identified as important to the realization of this vision is the implementation of AI.[18]

In brief, there is a large number of hurdles within medicine and health care that could potentially be overcome with the help of AI. However, alongside the great potential there are also significant social, ethical and legal challenges, especially with regards to the high stakes of life and death.[19] These challenges include risks for patient safety, treatment of outliers, concerns whether systems will be able to differentiate between

[13] J. Raymond Geis and others, 'Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement' (2019) 293 Radiology 436.

[14] Alejandro Rodriguez-Ruiz and others, 'Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists' (2019) 111 JNCI: Journal of the National Cancer Institute 916.

[15] E.g., Eric Racine, Wren Boehlen and Matthew Sample, 'Healthcare uses of artificial intelligence: Challenges and opportunities for growth' (2019) 32 Healthcare Management Forum 272, World Health Organization, *Ethics and governance of artificial intelligence for health: WHO guidance* (n. 11), K. H. Yu and I. S. Kohane, 'Framing the challenges of artificial intelligence in medicine' (2019) 28 BMJ Qual Saf 238.

[16] E.g., Minna Ruckenstein and Natasha Dow Schüll, 'The Datafication of Health' (2017) 46 Annual Review of Anthropology 261.

[17] E-hälsa 2025, 'Om vision e-hälsa 2025' (2021) <https://ehalsa2025.se/visionen/> accessed 2021-05-20.

[18] E-hälsomyndigheten, *Fokusrapport – Artificiell intelligens och e-hälsa*, (2020).

[19] E.g., A. Blasimme and E. Vayena, 'The Ethics of AI in Biomedical Research, Patient Care, and Public Health' in S. Das, Pasquale, F. and Dubber, Markus D. (ed.), *The Oxford Handbook of Ethics of AI* (Oxford University Press 2020), Titti Mattsson and Vilhelm Persson, 'E-hälsa' in Kavot Zillén, Titti Mattsson and Santa Slokenberga (eds.), *Medicinsk rätt* (Nordstedts Juridik 2020).

correlation and causality, and risks of unfair treatment due to bias regarding such as gender, ethnicity and age. Other risks are *competence loss* (in the sense that human knowledge of certain medical skills may erode, due to AI-systems taking over or heavily assisting the task), *automation bias* (if health professionals are over reliant towards AI systems), *overtreatment and overmedicalization*, *risk of integrity breaches*, as well as concerns about the *effects on responsibility, liability and trust*.[20]

Currently, and drawn to their extremes, at least two different discourses are demonstrated simultaneously; an idea of solutionism wherein AI-systems and methodologies will be the answer to if not all, at least most, problems, as well as a dystopian view of maleficent biased autonomous systems. Both views are adhering to deterministic views of technology, but, as laid out by Bucher: "there is nothing inherently neutral about algorithms or biased about humans, these descriptive markers emerge from particular contexts and practices."[21]

There are more balanced hopes for AI in health care, yet, the large interest, great expectations and inflated hopes seem to be fueled by the stakes involved: adopting AI in health care can literally be a matter of life and death. It also represents a possible profitable area of application for product developers, making commercial interest a contributing factor as well. However, as emphasized by legal and medical researchers, the use of AI in healthcare might also undermine traditional principles of medical law and patients' rights.[22] Another question that is raised is if patients have the right to refuse being subject to AI-systems.[23] But what is it with AI, compared to previously implemented technologies, that constitutes grounds for concerns?

[20] E.g., Obermeyer and Emanuel, 'Predicting the Future – Big Data, Machine Learning, and Clinical Medicine' (n. 12), Ziad Obermeyer and others, 'Dissecting racial bias in an algorithm used to manage the health of populations' (2019) 366 Science 447, Jessica Morley and others, 'The ethics of AI in health care: A mapping review' (2020) 260 Social Science & Medicine 113172.

[21] Taina Bucher, *If…then: algorithmic power and politics* (Oxford University Press 2018) 56.

[22] Iñigo de Miguel, Begoña Sanz and Guillermo Lazcoz, 'Machine learning in the EU health care context: exploring the ethical, legal and social issues' (2020) 23 Information, Communication & Society 1139.

[23] T. Ploug and S. Holm, 'The right to refuse diagnostics and treatment planning by artificial intelligence' (2020) 23 Med Health Care Philos 107.

1.2 The novelty of AI

Neither technical complexity nor information technologies are novel to Swedish health care, so what is new with AI-systems and methodologies in the clinical setting? Health professionals do not know the inner workings of all non-AI medical equipment, but hopefully the main logic behind their results or have a *human-in-the-loop* who could provide such explanations if needed.[24] The following characteristics of novelty, and associated opportunities and risks, are gathered from both the growing medical AI literature, as well as the wider media and communications and STS literature on automation and *datafication* in contemporary society. In short, from a transparency-focused and sociotechnical approach, AI-systems and methodologies may contribute to:

1. An increased *automation* of decision-making processes, with the benefits of efficiency, speed and avoiding dependency on overworked medical staff, which of course is highly attractive to these domains, but also the risks of reproducing historical skewness without sufficient oversight or scrutiny (including the impact of automation bias on human decision-making).[25]
2. *Large-scale adoption* as a result of automation. While having similar benefits as automation, it also entails both the advantage of excellence not being limited to certain human actors, as well as the heightened risk of errors or subjective prejudice decisions being applied on a large-scale as built-in features affecting large populations.[26]
3. *Opacity* resulting from the "black-box" nature of some AI-systems, with the risk of lacking explainability in complex algorithmic models, or systemic lack of transparency as AI-systems are applied in proprietary settings, with a complex array of data-sharing entities.[27]
4. *Data-dependency*, with large-scale quantification and "datafication" of everyday activities, which at best contributes to insights-driven incen-

[24] Jens Christian Bjerring and Jacob Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (2021) 34 Philosophy & Technology 349 364 P.364.

[25] C.f. Stefan Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (2019) 103 Droit et société 573.

[26] C.f. studies by media sociologist Jonas Andersson Schwarz, 'Platform Logic: An Interdisciplinary Approach to the Platform-Based Economy' (2017) 9 Policy & Internet 374; or on the platformisation of data-driven platforms, Stefan Larsson, 'Putting trust into antitrust? Competition policy and data-driven platforms' (2021) European Journal of Communication 02673231211028358.

[27] This is extensively developed in the following section.

tives and evidence-based decision-making, but with the risk of being at odds with established and regulated ideas of privacy, data-minimisation and rights to be forgotten, and that could result in the creation of what could be termed medical surveillance.[28]

5. *Obscured causability*. While medicine has always been considering multiple factors simultaneously (anamnesis, blood samples, measurements, etc.), automated medical decisions and classifications could propose hardships of deciphering which variables have led to a decision, and whether co-occurrences are wrongfully treated as causes.[29]
6. A *personalisation* of medicine, with the possibility of tailored drugs and treatments, as well as the risk of privacy breaches, challenged autonomy, mistreatment and discrimination.[30]
7. An increased *private-public complexity*, which is of particular relevance in Sweden, as Swedish health care is to a large extent public, while at the same time reliant on corporate service developers. The complex intertwinement of private and public dimensions are also relevant in terms of how this complexity can be handled from a regulatory perspective. It may be problematic from the public scrutiny point-of-view, if it hinders transparency of public sector organisations (see point 3 above,) or from the challenge of balancing the benefits of publicly collected data being used to train private AI-systems sold on markets.[31]

[28] E.g., Ruckenstein and Schüll, 'The Datafication of Health' (n. 16).

[29] E.g., Andreas Holzinger and others, 'Causability and explainability of artificial intelligence in medicine' (2019) 9 WIREs Data Mining and Knowledge Discovery e1312, Sendhil Mullainathan and Ziad Obermeyer, 'On the Inequity of Predicting A While Hoping for B' (2021) 111 AEA Papers and Proceedings 37.

[30] For an overview of both promises and pitfalls, see the editorial for a special issue on the subject, T. Feiler and others, 'Personalised Medicine: The Promise, the Hype and the Pitfalls' (2017) 23 New Bioeth 1.

[31] For example, pointed to as a challenge in studies on smart cities, Robert Brauneis & Ellen P. Goodman, 'Algorithmic transparency for the smart city' (2018) 20 *Yale JL & Tech.* 103, as well as pointed to in terms of the importance of improved procurement by the High-Level Expert Group on AI. See also Mattsson and Persson, 'E-hälsa' (n. 19) and Obermeyer and Emanuel, 'Predicting the Future – Big Data, Machine Learning, and Clinical Medicine' (n. 20).

2 Framing AI-transparency in health care

This section outlines main theoretical notions of AI-transparency in general, and in relation to medicine and health care in particular, in order to facilitate an analysis of its role for patients' rights. We argue that transparency is a wide concept, encompassing for example the more computer-scientific notion of explainability of AI-systems.[32] As mentioned, the lack of interpretability of AI is commonly described as "black-box".[33] The term can refer to a model that is too complicated to be interpretable, with only poor insights in how the training based on large data-sets reached a particular functionality or precision, or a model that is proprietary and hidden from external review.[34] It can also be both. This has led to a call for *transparency* of AI. Transparency is considered a key prerequisite for trustworthy AI, as stated by the EU high-level expert group and also mirrored in the current EU proposal for an AI regulation (AIA), published in April 2021.[35] But what does this idea of transparency entail?

2.1 AI transparency

Transparency is a multifaceted concept, as stressed by Larsson and Heintz,[36] often caught in a trade-off between different types of interests.[37] The term can be seen as a metaphor, where the material physical state of transparency – the see-through nature of an object – is deployed to describe cognitive, social, organizational phenomena and relations, and is

[32] E.g., Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).

[33] Frank Pasquale, *The black box society: the secret algorithms that control money and information* (Harvard University Press 2015), Brent Mittelstadt, Chris Russell and Sandra Wachter, 'Explaining Explanations in AI' (2019) Proceedings of the Conference on Fairness, Accountability, and Transparency 279.

[34] Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead' (2019) 1 Nature Machine Intelligence 206.

[35] High-Level Expert Group on Artificial Intelligence (HLEG), *Ethics guidelines for trustworthy AI*, 2019, hereinafter cited as HLEG 2019, Proposal for a Regulation of the European Parliament and of the council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 final. Hereinafter cited as AIA.

[36] Stefan Larsson and Fredrik Heintz, 'Transparency in artificial intelligence' (2020) 9 Internet Policy Review.

[37] For a discussion of seven different aspects, such as proprietary claims versus explainability versus human literacy, see Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).

based on the idea that what can be seen can be known (seeing is knowing).[38] Transparency can be considered a normative socio-legal ideal. In general, transparency is a broad concept, and this is also true regarding how the term is used in relation to AI. It can be used aiming at data underlying or used by AI models, algorithms and their logic, governance of AI-models, and so forth. Algorithmic transparency is a commonly used concept as well, but it could be misleading as AI in use is more than the function of algorithms,[39] and hence meaningful transparency need to encompass more than that.[40]

As pointed out above, there are several ways in which AI-systems can be opaque. Accordingly, three different forms of opacity are identified by Burrell: (1) intentional corporate or state opacity due to secrecy reasons, (2) opacity due to technical illiteracy and (3) opacity due to scale of operation of algorithms.[41] Another set of distinctions is made by Ferretti et al.: lack of disclosure, epistemic opacity and explanatory opacity.[42] Further, transparency in public decision-making can be described as information disclosure of different degrees, as described by de Fine Licht and de Fine Licht: informing about what the final decision (or recommendation or classification) is, about the process resulting in the decision (transparency in process) and about the reasons behind the decision (transparency in rational).[43]

Transparency is a vague concept, as pointed out by de Vries, stressing the need to ask: transparency of what, to whom, and when?[44] In addition, one must define what the problem is, if transparency is to be the answer. One issue, related to the vagueness, is the binary notion by which the concept of transparency is often used. Lee argues that we should not consider algorithms as binary, being either opaque or transparent. Instead,

[38] Larsson and Heintz, 'Transparency in artificial intelligence' (n. 36).

[39] Dignum, 'Introduction' (n. 3).

[40] Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).

[41] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) 3 Big Data & Society 2053951715622512.

[42] Agata Ferretti, Manuel Schneider and Alessandro Blasimme, 'Machine Learning in Medicine: Opening the New Data Protection Black Box' (2018) 4 European Data Protection Law Review (EDPL) 320.

[43] de Fine Licht and de Fine Licht, 'Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy' (2020) 35 AI & Society 917 918.

[44] Katja de Vries, 'Transparent Dreams (Are Made of This): Counterfactuals as Transparency Tools in ADM' (2021) 8 Critical Analysis of Law 121 124.

we need to consider them contextually and in practice, to analyse how agency and power is constructed. There are different degrees of agency and opacity in different parts of *algorithmic assemblages*, which are always situated in practice:

> ...algorithmic assemblages can be differently understood in different situations and are therefore neither completely opaque, nor completely transparent. Opacity is not only varying in degree or type, but also varies depending on the actor's situatedness.[45]

One important goal of transparency is to enable the assessment of, and demand for, fairness and accountability.[46] In the Swedish context, there is a far-reaching ideal of transparency of public administration in general. This encompasses publicly run health care and research institutions. Their decision making and procurement of technologies need to be to some degree interpretable, explainable and open for scrutiny. The patient's own access to electronic health data in journal systems is an example of transparency in practice, also pushed by legislation such as the GDPR.[47] The recently applied Medical Device Regulation promotes transparency within the health sector, for the purpose of medical safety.[48] In the AIA, transparency is also emphasized as a key component for safeguarding fundamental rights.[49]

2.2 Towards explainable AI

As a response to the call for AI-transparency, one part of the solution put forward is increased *explainability* (by some considered a concept included under the transparency 'umbrella').[50] The EU High-Level Expert Group on AI identify explainability as a core element of transparency, together with traceability and communication.[51] The urgency of transparency and explainability for AI in health care is echoed from a multi-

[45] Francis Lee, 'Enacting the Pandemic: Analyzing Agency, Opacity, and Power in Algorithmic Assemblages' (2021) 34 Science & Technology Studies 65 17.
[46] C.f. Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).
[47] GDPR (n. 5).
[48] MDR (n. 6).
[49] AIA, Explanatory Memorandum, 2.3 (n. 35).
[50] Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).
[51] HLEG 2019 (n. 35).

tude of perspectives.[52] Interpretability and explainability are considered crucial aspects to achieve trustworthiness, emphasized by for example the World Health Organization's guidance on ethics and governance of AI for health, according to which one of the key principles is to ensure transparency, explainability and intelligibility:

> AI technologies should be intelligible or understandable to developers, medical professionals, patients, users and regulators. Two broad approaches to intelligibility are to improve the transparency of AI technology and to make AI technology explainable.[53]

Explainable AI, also known as "xAI", can be defined as "a characteristic of an AI-driven system allowing a person to reconstruct why a certain AI came up with the presented predictions."[54] However, as with AI and transparency, there is not one agreed upon definition of what is included in the concept. Explainability has many facets and the terms transparency and interpretability are often used synonymously.[55] Lipton states that explainability is used to refer to some form of model interpretability, and distinguishes different forms that AI explanations can take: text (e.g., generated captions), visualizations (e.g., generated images), local explanations (e.g., gradient map masks, highlighting influential areas for classification of images) and explanation by example. The latter could be in the form of generated nearest neighbours and counterfactuals.[56] One could also distinguish between post-hoc explainable systems, providing local explanations on demand, and ante-hoc systems, built with "glass-

[52] E.g., Shaban-Nejad, Michalowski and Buckeridge, 'Explainability and Interpretability: Keys to Deep Medicine', Julia Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' (2020) 20 BMC Medical Informatics and Decision Making 310, Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24).

[53] World Health Organization, *Ethics and governance of artificial intelligence for health: WHO guidance*. Xiii (n. 11).

[54] Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' 2 (n. 52).

[55] Ibid. (n. 52).

[56] Zachary C. Lipton, 'The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery' (2018) 16 Queue 31, de Vries, 'Transparent Dreams (Are Made of This): Counterfactuals as Transparency Tools in ADM' (n. 44), Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31 Harvard Journal of Law & Technology (Harvard JOLT) 841.

box" approaches aiming to be interpretable by design.[57] Yet another distinction is whether the explanation should concern a system's general functionality, components of models, the training algorithm or rather of specific decisions.[58]

Insights on what constitutes a useful explanation can be found in diverse fields within the social sciences, as laid out by Miller, the main being that: why-questions are *contrastive* (responses to counterfactual cases), explanations are *selective*, causality is of greater importance than probabilities, and lastly, explanations are *social* as well as transfers of knowledge, in likeness to conversations or interactions.[59] Miller argues that all these factors converge around one single important point:

> [E]xplanations are not just the presentation of associations and causes (*causal attribution*), they are *contextual*. While an event may have many causes, often the explainee cares only about a small subset (relevant to the context), the explainer selects a subset of this subset (based on several different criteria), and explainer and explainee may interact and argue about this explanation.[60]

To be of use to involved parties, explanations should be "contrastive, selective, and social" rather than limited to models in science.[61] There is also an emphasis on "target audiences" in some of the literature on explainable AI,[62] that is, the awareness of that different types of addressees, such as medical staff, patients, and developers of AI-systems, will be having different types of need for what the explanations should hold.

On the other hand, the workings of AI-tools are also discussed as something that could deliberately be kept in the dark. To some extent this can be due to arguably valid reasons, such as protection of intellectual property rights or protection against maleficent gaming of systems or

[57] Holzinger and others, 'Causability and explainability of artificial intelligence in medicine' 5 (n.29).

[58] Mittelstadt, Russell and Wachter, 'Explaining Explanations in AI'. (n 33).

[59] Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences' (2019) 267 Artificial Intelligence 1.

[60] Ibid. p. 3.

[61] Mittelstadt, Russell and Wachter, 'Explaining Explanations in AI' (n. 33).

[62] E.g., Alejandro Barredo Arrieta and others, 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI' (2020) 58 Information Fusion 82.

security breaches.[63] Even though actors share the main goal of wanting to improve health care and well-being of patients, there can be reasons, intentional or unintentional, for actors to oppose meaningful transparency of their products. Bucher discusses how algorithms can function as *strategic unknowns*, where the opaqueness is deliberately maintained and could also be used as an advantage (which could, for example, be an inherent incentive in the private-public complexity pointed to above). Bucher argues that there is a risk of misplaced focus to keep telling the tale of the "black box"-ness of algorithms as something static and unavoidable, since it could serve different functions and be used as excuse for letting them continue to be kept opaque.[64]

It has been pointed out that there is a trade-off between better performance and explainability, meaning that improved prediction and accuracy, by applying more complex techniques, comes at the cost of decreased possibilities to interpret the models. However, Rudin argues that such a trade-off is not always given; sometimes there are models that score high on both interpretability and accuracy. Instead of trying to make black-box models explainable post-hoc, it should be an ex-ante concern to choose inherently interpretable models if they are going to be used for high-stakes areas such as health care.[65] In the medical context, relevant results can be found from diverse sets of data, hence it needs to be possible for practitioners to understand how and why a decision was made, as noted in the point on obscured causality in the novelty characteristics. Holzinger et al. argue that we need to go further than *explainable* AI, we also need *causability* to reach actual explainable medicine, by providing "causes of observed phenomena in a comprehensible manner through a linguistic description of its logical and causal relationships."[66]

Moreover, opaqueness of black-box medicine is at conflict with ideals of patient-centered medicine, argues Bjerring and Busch.[67] If AI-supported systems are expected to perform better than physicians, this creates a situation of epistemic obligation, where the physician has to follow

[63] C.f. Larsson, 'The Socio-Legal Relevance of Artificial Intelligence' (n. 25).

[64] Bucher, *If…then: algorithmic power and politics* (n. 21).

[65] Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead' (n. 34).

[66] Holzinger and others, 'Causability and explainability of artificial intelligence in medicine' (n. 29).

[67] Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24).

the system's recommendation. If not sufficiently explainable, they cannot explain how and why they give the recommendation or come to a certain conclusion. Hence, the patient cannot be adequately informed and not make an autonomous and rational decision.[68] What is needed is to make the AI *decision-making* explainable, which is required both by the patients and the physicians, for the latter not to be merely operators of non-comprehensible AI decisions.[69] If clinical decision support systems are omitting explainability, it threatens core ethical values of medicine, Amann et al. argue. From a legal perspective, they identify three core fields where xAI in health care is needed: "(1) Informed consent, (2) Certification and approval as medical devices (acc. to Food and Drug Administration/FDA and Medical Device Regulation/MDR) and (3) Liability."[70]

In sum, transparency and explainability can be considered important tools to even power-imbalances of the information (and knowledge) asymmetry at play in the health sector. Before an analysis of patient rights, and how they relate to AI, obligations of care providers and health professionals and information flows, we need to consider the state of rights and obligations in this context.

3 Rights, obligations and power dynamics in the context of health care

The following section outlines patients' rights of most relevance to transparency and explainability in relation to AI, primarily focusing on health care equality, issues of privacy and integrity, the right to information and patient autonomy, as well as what impact the implementation of AI could have on them.

There are national differences in how legal frameworks position the patient. For example, in Norway there is a patient's *rights* act,[71] while in

[68] Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24).

[69] Thomas Hoeren and Maurice Niehoff, 'Artificial Intelligence in Medical Diagnoses and the Right to Explanation' (2018) 4 European Data Protection Law Review (EDPL) 308.

[70] Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' p. 3 (n. 52).

[71] E.g., E. M. Aasen and B. M. Dahl, 'Construction of patients' position in Norway's Patients' Rights Act' (2019) 26 Nurs Ethics 2278.

Sweden the legal framework that specifically concerns health care is not expressed as a regulation of patients' legal rights, but in terms of *obligations* of actors (state, region, health care providers, health professionals) towards patients (in the Health and Medical Service Act, Patient Act, Patient Safety Act, Patient Data Act).[72] In this way the patient's rights are only implicitly expressed and a patient has limited possibilities to legally challenge medical decisions by judicial proceeding, as few of them are considered administrative legal decisions, argues Johnsson.[73] Health care providers and health professionals have to provide medical care in accordance to the legislation and can be held accountable for any wrongdoings. Rights in relation to health care can also be considered to belong to the public at large, with transparency as tool for accountability of public administration in accordance with the public access to information principle in Swedish law.[74]

In health care there are many actors involved: there are those who provide, those who receive, and those who facilitate or steer care. The different roles come with a difference in power. The starting point of discussing patients' *rights* and caregivers' *obligations* is in itself telling of a structure of power dynamics. The patient is considered in need of rights towards caregivers and the health system, due to the fact that the patient is considered to be in a position of less power in comparison to the other actors. Simultaneously, health professionals have obligations to not abuse their position of power. This dynamic can be found in several different instances, such as the patient being exposed to physical examinations and procedures, possibly life-saving or life-threatening, but also in terms of knowledge and information. The patient is in the hands of health care systems and practitioners who in general know more about how the system works, the specific medical condition and the treatments, as well as the medical state and personal sensitive health information of the individual patient (although this could be argued to not always be the case). This constitutes an *information asymmetry*. In the context of AI in the health sector, additional actors in this equation are the developers of AI models, tools and systems; that is, computer scientists, statisticians,

[72] Hälso- och sjukvårdslag (SFS 2017:30), Patientlag (SFS 2014:821), Patientsäkerhetslag (SFS 2010:659), Patientdatalag (SFS 2008:355).

[73] Lars-Åke Johnsson, 'Patientens ställning i vården och personalens skyldigheter' in Kavot Zillén, Mattsson, Titti, Slokenberga, Santa (ed.), *Medicinsk rätt* (Nordstedts Juridik 2020) 73.

[74] Tryckfrihetsförordning (1949:105), ch. 2.

medical researchers and more, and also the commercial entities or public suppliers of end products for screening, data handling or various types of predictions, etc. These other actors are in possession of yet another set of knowledge and informational power that health professionals and patients lack.

3.1 The right to (equal) health care

While access to health care varies greatly, globally as well as within nations, the right to health care is included as a basic human right, by article 25 of the United Nations' Universal Declaration of Human Rights.[75] In Swedish law, The Health and Medical Service Act reads that medical care should be provided with respect to equal value of all humans and the dignity of each individual.[76] In addition, the Declaration of Geneva points out; "I WILL NOT PERMIT considerations of age, disease or disability, creed, ethnic origin, gender, nationality, political affiliation, race, sexual orientation, social standing or any other factor to intervene between my duty and my patient."[77] Even though the said ideals and regulations exist, medicine and health care are not free from prejudicial and discriminatory practices. This is exemplified in reports by health professionals[78] as well as in research, such as by studies showing that women's expressions of pain are treated less serious than those of men,[79] and cases of racist interpretations leading to misdiagnoses or deprivation of treatment.[80] When AI-systems are trained on historical (or simply biased) data, mistreatment and discrimination could be reproduced and upscaled. This demands an awareness of what is built into the processes of data collection, labelling and interpretation, being the basis for learning algorithms.[81] Discrimi-

[75] Universal Declaration of Human Rights, (United Nations 2021), <https://www.un.org/en/about-us/universal-declaration-of-human-rights> Art. 25 accessed 2021-05-25.

[76] Hälso- och sjukvårdslag (SFS 2017:30).

[77] Declaration of Geneva (n. 1).

[78] Joakim Andersson, 'Läkare i stort upprop – vill se åtgärder mot rasism i vården' (2021) Läkartidningense.

[79] Anke Samulowitz and others, '"Brave Men" and "Emotional Women": A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain' (2018) 2018 Pain Research and Management 6358624.

[80] Sarah Hamed and others, 'Racism in European Health Care: Structural Violence and Beyond' (2020) 30 Qualitative Health Research 1662.

[81] E.g., Wiegand, T. et al. (ITU), *Whitepaper for the ITU/WHO Focus Group on Artificial Intelligence for Health* (The International Telecommunication Union 2018) 3–4.

natory practices could be reproduced due to AI-systems learning from history (status quo bias). For example, if a system is trained on a set of previously given treatments, or costs of previous treatments, groups that have a history of more easily receiving treatment could be incorrectly classified as higher risk patients.[82] This exemplifies the danger of treating covariances as explanations.

Besides equal care, The Swedish Health and Medical Service Act states that the person *most in need* should be prioritized,[83] which is challenged on the "quasi-market" of online doctors.[84] In theory, AI can help to find an answer to the conundrum of distinguishing who is most in need, by the ability to analyse large sets of data in a short amount of time and taking more variables into account. Well-trained algorithms could provide better predictions, finding previously unknown patterns or risks, for example in x-ray screenings or by improving prioritizing during triage by more accurate predictions of risk of re-admission or even death. However, the principle of most in need entails the necessity to be able to motivate decisions within health care – why one person is prioritized or not (for example during a triage process).

Further, the Swedish Patient Safety Act states that medical care should be conducted in *accordance with science and proven experience*.[85] Also, the MDR demands that evidence for clinical performance is provided.[86] A challenge of these principles is that it is ill-defined what should constitute the determinants of accuracy. AI models can also be working well for the large majority, but be less sensitive for identifying outliers and atypical symptoms or rare diagnoses, proposing the risk of patients being discriminated or mistreated, even when models on paper reach a standard of accuracy.

Transparency and explainability of implemented AI-systems are needed to be able to assess *fairness*, the equality of care, and that the persons most in need are in fact given priority and are treated in accordance with science and proven experience. Without proper information, these factors cannot be assessed.

[82] Obermeyer and others, 'Dissecting racial bias in an algorithm used to manage the health of populations' (n. 20).

[83] Hälso- och sjukvårdslag (SFS 2017:30), 3.1.

[84] Peter Bergwall, *Exploring Paths of Justice in the Digital Healthcare: A Socio-Legal Study of Swedish Online Doctors*, 51 (Faculty of Social Sciences, Lund University, 2021).

[85] Patientsäkerhetslag (SFS 2010:659), 6.1.

[86] MDR (n. 6).

3.2 The right to privacy and integrity

A motor in the development of AI models is the access to large amounts of reliable, accurate and representative data in order to train models and test their validity. A challenge for health care and medicine is the sensitive nature of the data needed.

A person's right to privacy is declared in the Universal Declaration of Human Rights, in the European Convention for Human Rights and in the EU Charter.[87] The right to respect for private life is also emphasized in medical ethics. The Declaration of Geneva states "I WILL RESPECT the secrets that are confided in me, even after the patient has died."[88] In recent time, important legal advances to strengthen individuals' right to privacy and control of personal information have been enforced in the form of the GDPR. According to Art. 9 GDPR, health data is a sensitive category of personal data, together with biometric and genetic data (when used for purpose of identification). In principle the processing of health data is prohibited unless there is an applicable exception that is listed in Art. 9.2 GDPR. The most relevant exceptions for processing sensitive data in health care are:

- after explicit consent (Art. 9.2(a)),
- if necessary for the protection of vital interests of a data subject incapable of giving consent (Art. 9.2(c)), for example in the case of an unconscious patient that needs treatment,
- if necessary for the assessment of a medical diagnosis, provision of health care or management of health care systems and services, if the data are processed by an actor legally bound by professional secrecy and confidentiality (Art. 9.2(h)),
- if necessary for reasons of public interest in the area of public health (Art. 9.2(i)) or for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes under certain provisions (Art. 9.2(j)).[89]

[87] Universal Declaration of Human Rights, Art. 12 (no arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation), European Convention for Human Rights, Art. 8 (Respect for private and family life), and the Charter of Fundamental Rights of the EU (2000/C 364/01), Art. 7 (Respect for private and family life) and Art. 8 (Protection of personal data).

[88] Declaration of Geneva (n. 1).

[89] GDPR, Art. 9, Art. 9.2(a), Art. 9.2(c), Art. 9.2(h), (n. 5).

The last exception represents a wider interest than that of the individual patient, and is therefore also of interest in relation to the public-private complexity accounted for in section 1.2. How should the line be drawn for the public interest exception when publicly held patient data are used for the training of private companies proprietary AI-systems? How can the public interest be ensured when the data is utilised by private interests, albeit for systems used in public health care?

According to the Swedish Public Information and Secrecy Act, health professionals have a legal obligation of confidentiality regarding patients' health conditions and other personal information, if the information cannot be shared without the individual or their relatives suffering.[90] Also, The Patient Safety Act states that a person working (past or presently) within health care is not allowed to share any information regarding an individual's health or other personal conditions, obtained in course of the work.[91] The Swedish Patient Act and Patient Data Act state that personal information should be registered and further processed with respect to the integrity of patients and others.[92]

The purpose for which sensitive information can be processed within health care is broad. The Swedish Act of complementary provisions to the GDPR, states that processing of sensitive personal information in health care is allowed, if necessary, for reasons such as preventative health care and medicine, medical diagnosis, providing health care or treatment, administration of health services and systems.[93] Further, processing of sensitive information for statistical use is permitted when benefits clearly outweigh potential risks for the privacy of individuals.[94] The Nordic countries' national registers could function as "goldmines" of health data for training algorithms. According to the Swedish Patient Data Act, processing of personal (health) information is permitted for national and regional registers, if consent is given.[95] Data from different registers are also combined to perform research and improve health care and medical knowledge. To facilitate longitudal studies, data need to be able to tie to the same individual to follow how health evolves over time. This is also

[90] Offentlighets- och sekretesslag (SFS 2009:400), ch. 25.1.

[91] Patientsäkerhetslag (SFS 2010:659), ch. 6.12.

[92] Patientlag (SFS 2014:821), ch. 10 and Patientdatalag (SFS 2008:355), ch. 1.2.

[93] Lag (2018:218) med kompletterande bestämmelser till EU:s dataskyddsförordning, 3.5.

[94] Lag (2018:218) med kompletterande bestämmelser till EU:s dataskyddsförordning, 3.7.

[95] Patientdatalag (SFS 2008:355), ch. 7.

true for research on for example how social, economic and demographic factors affect health.

Access to large sets of health data is crucial to develop accurate and fair AI models.[96] By the use of natural language processing, patient journals can also be important information sources by which vital knowledge could be gained. However, one problematic aspect of using patient journals is that less structured and standardized data provide challenges in the control of what could be revealed in the process of data sharing and learning of algorithms.

Here we have identified an important double bind: while big sets of health data constitute a necessity for AI development, data sharing – for example between public entities and commercial product developers – can pose challenges. Even though stripped from personal identifications, there is a risk of back door identification by reconstruction of aggregated data and combining of data sources. The principles of privacy and doctor's confidentiality could be contested by the data hunger of AI development and the increasing diffusion of data flows.

3.3 The right to information

The High-Level Expert Group on AI states that explanations should be timely and adapted to the level of expertise of the receiver.[97] This is also in line with the demands of the Swedish Patient Act,[98] which specifies that the caregiver must provide the patient with information regarding, for example, their health condition, methods for examination, expected course of treatment, any risks for complications and side-effects and methods to prevent diseases or injuries. The same act stipulates that information needs be tailored to the receiver's age, maturity, language background, and other individual preconditions, and that the one providing the information should make sure, as far as possible, that the content and significance of it has been understood.[99] Complaints by patients should also be answered with the receiver's ability to obtain the information in mind, and the health provider is obliged to provide an *explanation* of the

[96] Wiegand, T. et al (ITU) *Whitepaper for the ITU/WHO Focus Group on Artificial Intelligence for Health* (n. 81) 3.

[97] HLEG 2019 (n. 35).

[98] Patientlag (SFS 2014:821), 3.1.

[99] Patientlag (SFS 2014:821), 3.6, 3.7.

course of events, and describe actions planned for a similar event not to occur again, as stated by the Patient Safety Act.[100]

In recital 43 of the Medical Device Regulation, transparency and access to information are emphasized as "essential in the public interest, to protect public health, to empower patients and health care professionals and to enable them to make informed decisions, to provide a sound basis for regulatory decision-making and to build confidence in the regulatory system."[101] It also stipulates that information should be "appropriately presented for the intended user."[102] In addition, the GDPR states that data subjects (in this context: the patients) have the right to access information on data treatment in a "concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child."[103] It also stipulates a requirement for ex ante notification that should contain information about the purpose of the processing, how long the data will be kept and by whom the data will be processed. In addition, data subjects have the right to access information that the data controller holds about them: what categories of information, as well as copies of data, purpose of processing and with whom it is shared.[104]

In general, re-use of data for another purpose is prohibited, unless the re-use is for a purpose that is compatible with the initial purpose, such as research or statistical analysis,[105] or if a data subject consents ("downstream consent") to further processing for a new, incompatible, purpose.[106] In either case of further processing, data subjects should be notified. However, if providing a notification directly to data subjects is considered impossible or a disproportionate effort, especially for uses for scientific or statistical purposes, there is the option of making information publicly available instead, providing a basis for the opt-out principle for research studies and register data use[107] The GDPR is also demand-

[100] Patientsäkerhetslag (SFS 2010:659). 3.8(b).

[101] MDR, Recital 43 (n. 6).

[102] MDR, Recital 43 (n. 6).

[103] GDPR, Art. 12 (n. 5).

[104] GDPR, Art. 15 (n. 5).

[105] GDPR, Art. 5.1(b) and GDPR Art. 6.4 (n. 5).

[106] GDPR Art. 6.1(a) and GDPR Art. 6.4 GDPR. (n. 5), Regarding downstream consent, see Article 29 Working Party *Opinion 15/2011 on the definition of consent*, 2011, WP 187, p. 19.

[107] GDPR, Art. 14.5 and Art. 89.1 (research exception) (n. 5).

ing notification of the existence of solely automated decision-making, including profiling, with "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[108] In addition, Recital 71 states that in the case of automated decision-making or profiling, data subjects should have the right to: "obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision."[109]

Legally, health care providers are obliged to provide information in a way that the patient can understand. This could present a great challenge even without opaque AI systems. Conversational AI could provide solutions and improvements regarding this, but also pose difficulties as the sensitive nature of information given could require empathic skills and experience. Regardless, as previously discussed, health professionals do have information obligations in contact with the patient, to fulfil the requirements stipulated in the Swedish health care legislation. The versatility in the need for information means that AI-supported systems in health care have to be explainable with diverse levels of specificity, in different stages of implementation, to be intelligible by different actors or audiences.[110] It could mean initially providing explanations suitable for the developers of the system themselves, then for health professionals and people responsible for certification and procurement, and further challenging, to all relevant patients. Medical ethics and legal framework call for meaningful information, by contextual transparency and explainability, for caregivers to be able to fulfil their obligations and cater to patients' rights to information.

3.4 The right to dignity and autonomy

A main principle of medical ethics is the autonomy of the patient, which in both the normative and legal frameworks is tied to the dignity of human beings. The Declaration of Geneva states "I WILL RESPECT the

[108] GDPR, Art. 13.2(f) (n. 5), which refers to Article 22, that also adds the provision that the decision-making has to have "legal effect". See also Section 3.4.

[109] GDPR, Recital 71 (n. 5).

[110] Patientlag (SFS 2014:821), 3.6, 3.7. Also see e.g., Barredo Arrieta and others, 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI' (n. 62).

autonomy and dignity of my patient."[111] Part of this principle is that no treatment should be performed without informed consent, if not impossible to obtain due to a medical condition. Health care should as far as possible be planned and conducted in consultation with the patient, with consideration and respect, as stipulated by The Swedish Patient Safety Act.[112] The patient's autonomy and integrity ought to be respected and before consent, proper information must be provided, according to the Swedish Patient Act. When there are multiple treatment options (in line with science and proven experience), the patient should be able to make an informed choice.[113]

The principle of autonomy and choice is highly intertwined with the previously discussed principle of right to information, since real autonomy and choice could hardly be achieved if the patient has not had access to underlying information, and hence no possibility to interpret it. The idea of informed consent and individuals' control over their personal information is a main part of the GDPR. Article 22 regulates decision-making based on *solely* automatic processing, giving individuals the right not to be subject to such processing, requiring consent.[114] This regards automated processing for decisions that have a *legal effect* (or similar significant effect – which is to be interpreted as including all medical decisions). However, it is not established how the word *solely* should be understood in the context of health care.[115] If AI-systems analyse and prepare decisions, which are merely confirmed by a human doctor, should this be considered solely automatic processing ?[116]

Another aspect of the dignity and autonomy of patients, with regards to AI use, is in the situation of communication and information exchange. Is it a prerequisite for the dignity of a patient to have a human present, to talk to and provide information in an empathic manner? A hope for AI in health care is that the automation of certain tasks will free time for health professionals to be able to increase (or at least not reduce) the time spent with patients and on patient-close care. If realized, this is an opportunity to increase dignity in patient care and also autonomy,

[111] Declaration of Geneva (n. 1).

[112] Patientsäkerhetslag (2010:659), 6.1.

[113] Patientlag (SFS 2014:821), ch. 4, 5 and 7.

[114] GDPR, Art. 22 (n. 5).

[115] Hoeren and Niehoff, 'Artificial Intelligence in Medical Diagnoses and the Right to Explanation' (n. 69).

[116] Ibid. (n. 69).

if additional time can be allocated to information exchanges with patients and gaining knowledge of patients' preferences. However, parts of this could be (and are already) subject to automatization. Chatbots on caregivers' websites can increase access. Some people could also prefer talking to a bot – or robot[117] – rather than a human-being on topics of sensitive nature, possibly lowering thresholds.[118] While having potential benefits, automation of information tasks in health care could contest the principles of dignity and self-determination of patients.

Apart from the legal requirement in Art. 22 GDPR that decisions should not be based *solely* on automated processing, current legal and normative frameworks do not yet specify the role of the human-in-the-loop, and the question is whether the patient has a right to a human doctor.[119] Patients rely on clinicians being able to convey explanations in an accurate and understandable manner, improving the patient's agency in terms of risk assessment and informed choice.[120] If AI is not explainable, it could pose a challenge for health professionals to provide enough information on the reasons for classifications and proposed treatments, for patients to exercise their right to autonomy. Further, the personalisation of medicine could enhance autonomy but also lessen the experienced control of individuals. It could also be considered intrusive in practice, depending on the development and information and room for action provided to patients.

[117] See for example Maria Kyrarini and others, 'A Survey of Robots in Healthcare' (2021) 9 Technologies 8, and Laetitia Tanqueray, Tobiaz Paulsson, Mengyu Zhong, Stefan Larsson and Ginevra Castellano, 'Gender Fairness in Social Robotics: Exploring a Future Care of Peripartum Depression' In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Association for Computing Machinery, ACM, 2022).

[118] E.g., Bergwall, *Exploring Paths of Justice in the Digital Healthcare: A Socio-Legal Study of Swedish Online Doctors* (n. 84).

[119] For discussions regarding this, see for example Hoeren and Niehoff, 'Artificial Intelligence in Medical Diagnoses and the Right to Explanation' (n. 69), Fabrice Jotterand and Clara Bosco, 'Keeping the "Human in the Loop" in the Age of Artificial Intelligence' (2020) 26 Science and Engineering Ethics 2455 and Therese Enarsson, Lena Enqvist and Markus Naarttijärvi, 'Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts' (2021) Information & Communications Technology Law 1.

[120] As pointed out by Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' (n. 52) and Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24).

4 Discussion

The development of AI-supported tools for medicine and health care is rapidly evolving, and their broad adoption is imagined for the near future. However, the social, legal and ethical implications of AI use and automated decision-making/automated decision-support, could present severe challenges for successful and responsible implementation. In this chapter, we provide a brief overview of the novelties that AI in health care bring about in comparison to previous technologies, in order to point to key aspects of what this entails for current legal and normative (medical ethical) principles, especially with regards to transparency and explainability.

While acknowledging what could be gained by the adoption of AI, we must also consider what could be disrupted. This urges us to look for frictions between the basic principles of the normative and legal frameworks of health care and the implementation of AI. As stated by Obermeyer and Emanuel: "this challenge will create winners and losers in medicine. But we are optimistic that patients, whose lives and medical histories shape the algorithms, will emerge as the biggest winners as machine learning transforms clinical medicine."[121] If their optimism is to be realized, the rights of patients cannot be overlooked. Equality of care, privacy, access to information, autonomy and dignity are basic principles of the legal framework of patients' rights. When developing and implementing AI-systems and methodologies to be used in the context of health care, there is indeed a need to jointly address how they could be compliant with these basic principles, and hopefully even support and strengthen them.

We argue that transparency needs to be understood as *situated* in the *information practices* of health care, in line with Lee's notion of algorithms in practice,[122] and not as a binary state of full transparency or opaqueness. Data flows in health care are based on medical ethical ideals and are two-faced; both carefully protecting and carefully providing information, between patients and the healthcare systems, as well as developers, registers, and other actors and infrastructures in public and private sector. This becomes evident in the right to privacy and the right to information,

[121] Obermeyer and Emanuel, 'Predicting the Future - Big Data, Machine Learning, and Clinical Medicine' 1218 (n. 20).

[122] Lee, 'Enacting the Pandemic: Analyzing Agency, Opacity, and Power in Algorithmic Assemblages' (n. 45).

which also constitute a foundation for the right to autonomy and dignity. The data flows represent both opportunity and risk, further emphasized by the *private-public complexity*. As stated above, this may be detrimental to transparency-requirements in assessment of whether the specific AI-tools or services actually work in a fair and trustworthy way, given proprietary interests or the need for keeping business secrets on competitive markets. This also stresses another balancing question of principal interest. On the one hand, it is feasible that market-driven players may indeed be best suited for developing new AI-systems to be utilised as products or services procured by the public sector. But, on the other hand, if that development is dependent on data collected in the public sector, from its patients, protected by the GDPR but shared in the name of the public interest, there may indeed also be a need for more thought on how to ensure that these applications actually serve the public interest, while under clear private interest custody.

Furthermore, if we are to benefit the most of AI in medicine, it is not to be used in a one-size-fits-all-manner, but to make the best decision out of all available information about the individual patient. The level of risk, or best treatment option, may depend on aspects such as age, gender or ethnic origin. Not taking these factors into account when applicable, could lead to discriminatory results by ignoring known (or unknown) risk factors and posing disadvantages to vulnerable groups. This *personalisation* could enhance, as well as contest, values of equal care and most in need, while at the same time pose risks for discriminatory bias of systems, privacy breaches and weakened autonomy, especially pushed by *automation* and *large-scale application*.

Transparency and explainability constitute prerequisites for assessment of patient's rights, demanding fairness and accountability. However, this leads to vital questions connected to the theoretical foundation of what type of transparency is to be aspired to, when, and to whom. Only asking for general transparency and explainability might not be meaningful unless further specified, as pointed out by de Vries.[123] Explainability could function as a tool for patient autonomy as well as sound scepticism and scrutiny, mending over-reliance on algorithms and algorithmic bias. If AI-systems are not sufficiently explainable, they could end up not being used by health professionals due to lack of trust, maybe depriving the

[123] de Vries, 'Transparent Dreams (Are Made of This): Counterfactuals as Transparency Tools in ADM' (n. 44).

person most in need of being prioritized, i.e., affecting the most in need principle (discussed in section 3.1). Transparency and explainability are necessary tools to assess how well AI-systems perform and align with scientific knowledge in the specific context of its use, calling also for *causability*, as Holzinger et al. suggest.[124] If the instrumental goal of transparency is increased knowledge – for health professionals, patients, and the public – information and explanations need to be adjusted accordingly, and be specific enough to be meaningful in individual cases. This could promote the goals of fostering agency, accountability and fairness, and in the long-run; trust. For the patient wanting to know the reason for being sent home from the ER, one cannot refer merely to an algorithm suggesting that it was the right call to make. Transparency needs to be *contextually* understood and, as Miller and Mittelstadt et al. point out, explanations are executed as social, selective and contrastive functions[125] – what is the smallest difference that would have resulted in a different decision, that is, that the patient is *not* sent home from the ER?

Still, one could ask, what should lead the way forward – best possible care or most transparency? While this is perhaps in some cases a false dichotomy,[126] the concept of best possible care could also be argued to not consist only of the physical health outcome. It also encompasses the adherence to ethical core values of how health care should be conducted.[127] There is also an epistemic aspect: how would we know if it is the best possible care, or even an accurate improvement, unless it could be assessed by sufficient transparency? Medical ethics and the legal framework do not allow for AI-tools not being transparent and explainable *in meaningful ways*, with careful consideration of the needs of different addressees. Such opacity would hinder the possibilities for health practitioners to interpret and assess a recommendation, decision or classification, and by that limit the possibilities for patients to get the information they are entitled to, to

[124] Holzinger and others, 'Causability and explainability of artificial intelligence in medicine' (n. 29).

[125] Miller, 'Explanation in artificial intelligence: Insights from the social sciences', (n. 59), Mittelstadt, Russell and Wachter, 'Explaining Explanations in AI' (n. 33).

[126] See the discussion on possible trade-off between accuracy and explainability in section 2.2.

[127] In line with the reasoning of Amann and others, 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective' (n 52), Bjerring and Busch, 'Artificial Intelligence and Patient-Centered Decision-Making' (n. 24) and Hoeren and Niehoff, 'Artificial Intelligence in Medical Diagnoses and the Right to Explanation' (n. 69).

make a correct risk assessment. AI-systems cannot be allowed to function as *strategic unknowns*,[128] neither in the form of decision-support nor in automated decision-making, in the context of health care. In the implementation of new technology, health professionals need to be given a chance to live up to medical ideals and regulatory requirements, because, no matter how accurate AI-technologies ever get, and how well they ever learn to imitate human compassion, they will never feel the burden of the Hippocratic oath.

5 Conclusions

In this chapter, we ask what role transparency and explainability of AI could have, in relation to patients' rights and information flows in Swedish health care. We outlined a set of novelty characteristics associated with AI-systems in health, including: automation, scale, opacity, data-dependency, obscured causality, personalisation and a private-public complexity. By first setting a foundation of a conceptual framework on transparency in general and explainability in particular, in relation to AI in health care, we analyse the legal and normative regulatory framework of patients' rights. We address those rights that are most relevant to transparency and explainability in relation to AI; the right to equal health care, privacy, information, dignity and autonomy, with the purpose of pinpointing main challenges in the implementation of AI in, primarily public, health care.

We find that it is not possible to adhere to the basic principles of the normative and legal frameworks of Swedish health care, if meaningful and contextual transparency and explainability are not deployed in the implementation of AI. Instead of focusing the highest quality of health care as something which stands on opposite side of the requirement for transparency (by the accuracy versus explainability trade-off), we argue for the need to consider them as interdependent. The best possible health care cannot be achieved without transparency. As transparency is situated in information practices, and not a binary state, the way forward is to find what kind of transparency that is needed to safeguard best possible health care. Meaningful and contextual transparency and explainability are necessary for the provision of patient autonomy and as a means to assess if the best possible care is given to the ones most in need.

[128] Bucher, *If…then: algorithmic power and politics* (n. 21). See section 2.2.

Katarina Fast Lappalainen

Protecting Children from Maltreatment with the Help of Artificial Intelligence: A Promise or a Threat to Children's Rights?

1 Introduction

In late 2020, a scandal was revealed that sent shockwaves throughout Sweden. A man, 40 years old, was reportedly found in a filthy apartment at the outskirts of Stockholm.[1] Bruised all over his body and in need of immediate care, he was malnourished, had no teeth and had difficulties expressing himself verbally. It turned out that he had not attended school since the age of 12 and had ever since been kept in isolation by his mother. Nevertheless, during the remainder of his childhood the boy remained enrolled at his school and continued to appear on the class lists and received incomplete grades, even though he was not attending school. His older sister, no longer living with the family, had notified the social services of the possible maltreatment of her brother. Despite

[1] Special thanks to the editors Mattias Dahlberg and Katja De Vries for invaluable comments and suggestions. I am also grateful to Paul Lappalainen, Swedish and American lawyer for important comments and language review. Finally, I wish to thank Sara Skoglund, social worker, systemic family therapist and court expert, U.K. and Pia Molander Wistam, Head of the Department of Children and Youth, Norrtälje Municipality, Sweden for valuable information and materials.

this, neither the school, nor the authorities acted upon the information available to them.[2]

Could this and other scandals where the community fails to protect children from maltreatment, have been prevented with the help of predictive tools using artificial intelligence (AI) for effective risk management?

Predictive tools for child protection based on AI have, with varying success, been developed in different parts of the world. Some examples are the *Vulnerable Children Predictive Risk Model* of New Zealand from 2012, the *Allegheny Family Screening Tool* used by Allegheny County in Pennsylvania in the U.S. since 2014 and the *Early Help Profiling System*, developed by Hackney County Council in London together with Xantura, a private company. In Scandinavia the *Gladsaxe-model* from Copenhagen, Denmark, seems to have been the first in the region as it was ready in 2018. In Sweden, the municipality of Norrtälje launched an AI tool to analyse cases based on preliminary warning referrals in 2020, to help in the detection of future cases of child maltreatment.[3]

It could be argued that such tools in general will help prevent maltreatment of children and enable social services to become more effective in their outreach work and thus the provision of support to children at high risk at a lower cost. The struggle to provide more effective child welfare and the reality of substantial funding cuts, common to the authorities in many European countries, increases the interest in such systems.

Contrary to the promise and hopes for AI tools is the fact that the use of such tools for child protection, comes with multiple risks from a children's rights perspective. This is certainly the case regarding the use of predictive risk modelling (PRM) in child welfare.

PRM can be developed using different techniques which can have somewhat different outcomes from a legal and ethical perspective. These techniques are described further in section 3. Models generally rely on a

[2] The case was reported in several media outlets, e.g., Sveriges Radio, https://sverigesradio.se/artikel/7613921 and Expressen 2020-12-01, *Släktingen hittade instängde mannen: det luktade ruttet* by Erik Wiman. In the end the mother was released as the prosecutor reportedly did not find evidence of any crime, since the son was not physically restrained from leaving the home. See e.g. *Åklagaren om instängde sonen: inga bevis för brott* by Niklas Eriksson, https://www.aftonbladet.se/nyheter/a/x3J3ln/aklagaren-om-instangde-sonen-inga-bevis-for-brott.

[3] Norrtälje Municipality website: https://www.norrtalje.se/ai_oro.

formalization of existing professional or actuarial expertise.[4] The variables in a model could, for example, be derived from the experience of child welfare practitioners who consider poor housing or single parenthood to be risk factors. Nevertheless, models based on former human experience or decision-making can be flawed, biased and out-dated.

A model can also be based on actuarial expertise showing the existence of a certain statistical relationship between a variable and child maltreatment, such as the personal history of child abuse of a care giver indicating a known statistical risk factor.[5] This means that the use of AI can provide result that are somewhat lacking in precision. AI models based on PRM,[6] meaning predictions based on as many variables as possible, even those that might seem irrelevant, are likely to be of limited precision and lacking context. A recent scientific mass collaboration, "The Fragile Families Challenge", shows that machine learning models are not very accurate when it comes to predicting life trajectories, which give reason to question their usefulness in the context of child welfare.[7]

Consequently, these models have the potential to lead to wrongful decisions, resulting in interventions that should not have taken place (false positives), as well wrongful decisions leading to a failure to act (false negatives).[8] The efficacy of such models can therefore be questioned from a methodological point of view.[9]

[4] Bosk, E.A. *What counts? Quantification, worker judgment and divergence in child welfare decision making*, Human Service Organizations: Management, Leadership & Governance, 2018, 42:2, p. 205.

[5] Bosk 2018, p. 214.

[6] Cuccaro-Alamin, Foust, Vaithianathan, Putnam-Hornstein 2017 p. 293.

[7] Salganik M.J. et al. Measuring the predictability of life outcomes with a scientific mass collaboration, Proceedings of the National Academy of Sciences of the United States of America (PNAS), April 2020 117 (15) 8398-8403, 2020 March 30, https://doi.org/10.1073/pnas.1915006117.

[8] V. Eubanks, *Automating Inequality – How High-Tech Tools Profile, Police and Punish the Poor*, Picador, New York 2019, p. 157; Bosk 2018, p. 205; S. Cuccaro-Alamin, R. Foust, R. Vaithianathan, E. Putnam-Hornstein, *Risk assessment and decision making in Child protective services: Predictive risk modeling in context*, Children and Youth Services Review, 2017, p. 292, p. 295, see https://www.datanetwork.org/wp-content/uploads/PRM-CYSR-article.pdf.

[9] R. van Brakel, *The Rise of Pre-emptive Surveillance: Unintended Social and Ethical Consequences*, Chapter 14 in E.R. Taylor and T. Rooney, 2016, Surveillance Futures: Social and Ethical Implications of New Technologies on Children and Young People, Routledge, London, p. 194–196.

Predictive models based on actuarial or human expertise can increase the risk for unlawful or disproportionate interferences in several of the rights of the child, such as the respect for family life, prohibition of discrimination, and protection of personal data. They can also affect the positive obligations of the state to protect children from torture or inhuman and degrading treatment or even child fatality if an AI system fails to detect children at a high risk of maltreatment. Moreover, the lawfulness regarding the possible use of variables such as socioeconomic status, health status and cultural or religious background of the parents needs to be addressed.[10]

Risk estimation may also be carried out through text mining, such as natural language processing topic modelling (NLP/TP). The latter is used to analyse texts and do not rely on risk factors that are top-down listed by a human expert. Instead, in such models the relevant variables are identified bottom-up by algorithmically analysing earlier cases which might give a more objective outlook if factors such as poor housing and single parenthood indeed are risk factors.[11] Nevertheless, these AI models also risk entrenching bias related to historical data,[12] which has been depicted by Medvedeva et al. as "status quo bias"[13]: for example, if child care services are more inclined to investigate cases of child maltreatment in families with poor housing and single parents, the data model will learn that these variables are relevant whereas in fact the training data might be misleading as they do not include undetected cases of child abuse in wealthier households with two parents.

The purpose of this paper is to give a preliminary overview and analysis regarding the design and use of AI tools to identify children at high risk of maltreatment in relation to relevant children's rights. Are such child

[10] See e.g. G. Van Bueren, *Opening Pandora's Box – Protecting Children Against Torture, Cruel, Inhuman and Degrading Treatment and Punishment* in G. Van Bueren (ed.), Childhood Abused – Protecting Children against Torture, Cruel, Inhuman and Degrading Treatment and Punishment, Routledge (e-book) 2018, p. 85; J. Ennew, *Shame and Physical Pain: Cultural Relativity, Children, Torture and Punishment* in van Bueren 2018 p. 53.

[11] Harrison C.J. and Side-Gibbons C.J., *Machine learning in medicine: a practical introduction to natural language processing*, BMC Medical Research Methodology, 2021, 21:158, p. 3.

[12] L. Svensson, *Automatisering – till nytta eller fördärv?* (Automation – to benefit or ruin?) Socialvetenskaplig tidskrift 2019:3-4, p. 358–359.

[13] M. Medvedeva et al., *The Danger of Reverse-Engineering of Automated Judicial Decision-Making Systems*, ArXiv 18 December 2020, p. 4, https://arxiv.org/pdf/2012.10301v1.pdf.

protection tools aligned with children's rights as laid down in the UN Convention of the Rights of the Child (UNCRC), the European Convention of Human Rights and the EU Charter of Fundamental Rights? And to what extent?

An analysis of law in relation to government use of AI tools for child protection needs to be undertaken from different perspectives. Applying children's rights requires a child-centred approach, which takes its starting point in the idea that children are equal as human beings and independent rights holders. Child maltreatment is a complex and multifaceted problem as is decision-making that relates to it. The perspective applied here is therefore holistic and interdisciplinary, meaning that the law is one of many tools to make children's rights real.[14] To this end, sources regarding for example social work and computer engineering are therefore of vital importance.

More specifically, this analysis revolves around the use of technology for child protection and the interaction and interdependency of "rules and tools", which embodies consequences both for the child and for society, and adds to the holistic and interdisciplinary perspective of legal informatics, which operates at the intersection of law and information and communication technology (ICT).[15] An important part of legal informatics is furthermore to contribute to the design of emerging technologies, such as predictive tools based on AI for child protection, through the establishment of legal standards and frameworks based in law.[16]

The outline of this paper is the following. In sections 2 and 3 I discuss the significance of preventive measures regarding child maltreatment and the different models used to predict child maltreatment. In section 4 I assess them from a legal perspective (UNCRC and ECHR). In the fifth and final section I draw some preliminary conclusions about the use of predictive tools based on AI to prevent child maltreatment.

[14] M. Grahn Farley, *Barnkonventionen – en kommentar*, Studentlitteratur, Lund 2019, p. 13–14; Van Bueren 2018.

[15] S. Greenstein, *Elevating Legal Informatics in the Digital Age* in S. Pettersson (ed.) Digital Human Sciences: New Objects – New Approaches, Stockholm University Press (2021) p. 156; P. Seipel, *IT Law in the Framework of Legal Informatics*, Scandinavian Studies of Law (2004) vol. 47, p. 37 f.

[16] C. Magnusson Sjöberg, *Om rättsinformatik* in C. Magnusson Sjöberg (ed.) Rättsinformatik – Juridiken i det digitala informationssamhället, Studentlitteratur, Lund 2021, p. 21–22.

2 The significance of early intervention and the idea of child protection systems to prevent child maltreatment

Child maltreatment is a highly complex, multidimensional problem both at an individual and societal level as well as from a biological and psychological standpoint. In Art. 19 of the UNCRC it is defined as "all forms of physical or mental violence, injury or abuse, neglect or negligent treatment, maltreatment or exploitation, including sexual abuse". Child maltreatment is not only detrimental to the individual child but to society as a whole, and comes with astronomical costs. It is regarded as a major public health issue and is assessed to affect at least 55 million children in the WHO European region alone (53 countries).[17]

In the 1990s neurological research demonstrated that child maltreatment can cause permanent neurological effects on children's brains.[18] This means that child maltreatment not only can cause lasting physical damage as the result of abuse or neglect, but affects behaviour, emotional well-being, personal relationships and cognitive functions.[19]

These research findings were followed by a shift in social work towards early intervention and evidence-based practices.[20] Instead of focusing on reactive approaches, such as providing protection for children who may already have been maltreated, the goal is to prevent it from happening. A central part of this proactive approach is risk assessment, which can be performed in various ways, normally through "operator driven" clinical or actuarial assessments.[21] A more recent trend is the development of technological tools using PRM to identify children at high risk of

[17] D. Sethi, Y. Yon, N. Prakeh, T. Anderson, J. Huber, I. Rakovac & F. Meinck, *European status report on preventing child maltreatment*, World Health Organization, 2018, p. 3 f.; D. Glaser, *The effects of child maltreatment on the developing brain*, Medico-Legal Journal 2014, vol. 82(3), p. 98.

[18] D. Daro and A.C. Donnelly, *Reflections on Child Maltreatment Research and Practice: Consistent Challenges* in D. Daro A.C. Donnelly, L.A. Huang, B.J. Powell (eds.) Advances in Child Abuse Prevention Knowledge – The Perspective of New Leadership, 2015, Springer (e-book), p. 8.

[19] Glaser 2014, p. 97.

[20] D. Daro and A.C. Donnelly 2015, p. 8.

[21] H. Vannier Ducasse, *Predictive risk modelling and the mistaken equation of socio-economic disadvantage with risk of maltreatment*, British Journal of Social Work 2020, p. 2 f.

being maltreated.[22] However, using standardized risk assessment tools based on actuarial principles is not a new idea within the field of social work. Already in the 1980s, the prospect of using expert systems for decision-making regarding child interventions was proposed by Schoech et al., while also acknowledging that such a system "offers many legal and ethical challenges to the human service professions".[23]

3 Cases of artificial intelligence for child protection

Today several tools using artificial intelligence have been developed or are under development for predicting child maltreatment in various countries such as Denmark, the Netherlands, New Zealand, Sweden, the U.K. and the U.S. However, the success rate varies and many of the projects have been discontinued.[24] The focus of this paper is limited to some of the most well-documented, researched and debated tools, most of which are accounted for as being PRM tools based on principles of actuarial risk assessment,[25] with the exception of the Norrtälje NLP/TP model. Each of the different models however is used to build tools to predict the risk for child maltreatment.

Before the presentation moves on to artificial intelligence (AI) and notably PRM tools, as well as NLP/TP more specifically, for countering child maltreatment, the question is: what are they?

PRM is a form of "predictive modelling", which is a description for tools with the aim of making accurate predictions, such as "machine learning", "AI" and "data mining". Kuhn and Johnson define predictive

[22] P. Gillingham, *Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the 'Black Box' of Machine Learning*, British Journal of Social Work, 2016, 46, p. 1045.

[23] D. Schoech PhD, H. Jennings, L. L. Schkade PhD & C. Hooper-Russell,1985, *Expert Systems – Artificial Intelligence for Professional Decisions*, Computers in Human Services, 1:1, 81–115, DOI: 10.1300/J407v01n01_06, p. 106.

[24] A. Møller Jørgensen, C. Webb, E. Keddel, N. Ballantyne, *Three roads to Rome? Comparative policy analysis of predictive tools in child protection services in Aotearoa New Zealand, England, & Denmark*, Nordic Social Work Research 2021, p. 2, https://doi.org/10.1080/2156857X.2021.1999846; P. Gillingham, *Decision Support Systems, Social Justice and Algorithmic Accountability in Social Work: A New Challenge*, Practice: Social Work in Action, 2019, Vol. 31, No. 4, p. 278.

[25] Gillingham 2016, p. 1045.

modelling as "the process of developing a mathematical tool or model that generates an accurate prediction".[26]

PRM, more specifically, can be defined as:

> a type of predictive analytics… a statistical method of identifying characteristics that risk-stratify individuals in a population based on the likelihood each individual will experience a specific outcome or event. The result of the model's mathematical algorithm is a risk score. Unlike model-building techniques traditionally used in risk assessment – in which variables are chosen on the basis of previously researched relationships with the specified outcome – in PRM, as many data points as possible are examined, even if there is no previously specified relationship with the outcome of interest.[27]

The model works through algorithms, i.e. an instruction to the computer with a series of steps or procedures to follow. The algorithms will be coupled with variables to create a mathematical model (machine learning).[28] The model can examine and learn from a large amount of data from a variety of sources such as administrative datasets, whereby hidden patterns, correlations, regularities etc. can be extracted, which in turn can help in making predictions of different kinds, such as predictions concerning future risks in fields as diverse as finance, health and meteorology.[29]

The result of the PRM tool is consequently a risk score that can be used to for example to support decision-making in child welfare.[30]

NLP/TP models can also be used for risk prevention. It can be described as a form of text-mining, which basically means that "a group of algorithms, reveal, discover and annotate thematic structure in a collection of documents". It has been used or tested for example in healthcare to predict disease risk, risk of hospital readmission or suicide.[31]

[26] M. Kuhn & K. Johnson, *Applied Predictive Modeling*, Springer, New York 2013, p. 1.

[27] Cuccaro-Alamin, Foust, Vaithianathan, Putnam-Hornstein 2017 p. 293.

[28] M. Broussard, *Artificial Unintelligence – How Computers Misunderstand the World*, MIT Press, Cambridge Massachusets 2018), p. 94.

[29] S. T. McKinlay, *Evidence, Explanation and Predictive Data Modelling*, Philosophy and Technology (2017) vol. 30, p. 462–464; S. Greenstein, *Our Humanity Exposed – Predictive Modelling in a Legal Context*, Stockholm University 2017, p. 22, p. 70 ff.

[30] Cuccaro-Alamin et al. 2017, p. 293 f.

[31] P. Kherwa and P. Bansal, *Topic Modeling: A comprehensive review*, EAI Endorsed Transactions on Scalable Information Systems 2019, p. 2 and 10; A. Rumshisky, M. Ghassemi, T. Naumann, P. Szolovits, V.M. Castro, T.H. McCoy and RH Perlis, *Predicting early psychiatric readmission with natural language processing of narrative discharge summaries*, Translational Psychiatry 2016, 6e921, doi:10.1038/tp.2015.182.

The use of these tools and techniques also presents certain challenges. First of all, the models are created by humans, which can be reflected in the design of a model. This can be both a strength and a weakness. The developers can be endowed with expert-knowledge regarding children at risk of maltreatment as well as experience and empathy. However, it also means that the developers of the tool have the potential to incorporate bias.[32] The developers might lack the necessary insight regarding the prejudices that can impact child welfare decision-making, such as the example of social workers that might be less likely to detect child maltreatment in wealthy, two-parent households.

These tools are also limited to what algorithms can actually process, including the availability of relevant data. The amount, quality and nature of data can be imperfect and incomplete, which can especially be the case regarding data concerning human behaviour.[33] Moreover, the processing of data within the tool is often referred to as a "black box", since it, unlike a human professional or expert, cannot provide any reasons for its predictions, meaning a lack of transparency.[34]

Another concern is that no such tool can be 100 percent accurate, which may result in results that are wrong. As stated by O'Neil:

> There would always be mistakes, however, because models are, by their very nature, simplifications. No model can include all of the real world's complexity or the nuance of human communication. Inevitably, some important information gets left out.[35]

The concern regarding accuracy therefore evokes important issues related to evidence and substantiation.[36] It can be discussed if assessments made by an AI tool should be used as evidence,[37] and more specifically what the probative value would be in a legal setting. Finally, the use of PRM and NLP/TP tools can raise various ethical and legal challenges, such as

[32] Broussard 2018, p. 67. O'Neil breaks down predictive modelling to the individual level and concludes that racism can be apprehended as a predictive model "whirring away in billions of human minds around the world. It is built from faulty, incomplete, or generalized data." See C. O'Neil, *Weapons of Math Destruction – How Big Data Increases Inequality and Threatens Democracy*, Penguin Books, USA, 2016, p. 22.

[33] McKinlay 2017, p. 463.

[34] Greenstein 2017, p. 73.

[35] O'Neil 2016, p. 20.

[36] Mcinlay 2017; Gillingham 2016, p. 1049–1052.

[37] McKinlay, p. 471–473.

racial discrimination and poverty profiling.[38] We can therefore not be certain that such tools will actually and effectively counteract child maltreatment.

3.1 The Vulnerable Children PRM

An initiative by the government of New Zealand in 2011 appears to be the first initiative in the world by a government to develop a PRM tool to predict child maltreatment, the Vulnerable Children PRM.[39] The initiative was part of a large-scale reform in child protection services, with a social investment approach,[40] which among other things included new legislation and the linking up of databases across public service systems.[41]

A team of researchers in economy, social work and ethics at the Centre for Applied Research in Economics (CARE) at the University of Auckland, New Zealand, was given the task of researching the question of whether it would be possible to use administrative data to identify children at high risk of maltreatment.[42] The team developed an algorithm drawing from a data set from public welfare benefit systems and child protection services. The children included in the analysis were children 1) identified with a family that had a benefit period, i.e., the length of time during which a family received some kind of social benefit between the child's birth and 2nd birthday, including pre-birth and pregnancy related periods and 2) born between January 2003 and June 2006, so that they would reach 5 years of age by the end of the sample period.[43]

The model made use of 132 predictor variables which were presented in five categories in the CARE report. The first two categories included

[38] Eubanks 2018, p. 158; Cuccaro-Alamin et al. 2017, p. 295.

[39] N. Ballantyne, *The ethics and politics of human service technology: the case of predictive risk modelling in New Zealand's child protection system*, Hong Kong Journal of Social Work, vol. 53, 2019, p. 15.

[40] Møller Jørgensen et al. 2021, p. 3; Ballantyne 2019, p. 18.

[41] Gillingham 2016, p. 1046.

[42] Ibid.

[43] CARE (2012), R. Vaithianathan, T. Maloney, N. Jiang, I. De Haan, C. Dale, E. Putnam-Hornstein, T. Dare, *Vulnerable Children: Can Administrative Data Be Used to Identify Children at Risk of Adverse Outcomes?* Centre for Applied Research in Economics, University of Auckland, New Zealand, p. 10 available at: https://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/research/vulnerable-children/auckland-university-can-administrative-data-be-used-to-identify-children-at-risk-of-adverse-outcome.pdf.

variables related to the care, protection and benefit of the subject child and that of other children in the family. Some examples are findings of abuse and neglect, child protection notifications, court orders and proportion of time on a benefit. The third and fourth categories consisted of data relative to characteristics concerning the child's caregiver and the family at the start of the period. For example, the data included gender, age, level of education, whether the household consisted of single or dual caregivers, number of children, age of caregivers when the oldest and the subject child were born etc. The fifth and final category concerned the care and protection and benefits history of the subject child's caregivers before the age of 16 as well as benefit histories in adulthood.[44]

It was determined that the model could accurately predict maltreatment within an area under the receiver operating characteristic (ROC) curve of 76 percent (a performance measurement for classification problems) which is comparable to the rate found in digital mammography.[45] The team also outlined a "business case" discussing return on investment in the PRM tool, which would mean a great reduction of the costs per child.[46]

The ethical approach taken by the team has been described as consequentialist.[47] In sum, the conclusion of the ethical evaluation was that the PRM tool certainly gave rise to concerns regarding certain aspects such as the risk of false positives, the fact that non-beneficiaries are not risk assessed and privacy issues etc.[48] As long as these concerns could be significantly mitigated or ameliorated, they could be outweighed by the important potential benefits of the tool.[49]

When the Vulnerable Children PRM became known to the public, it met with great concern. The accuracy of the tool was questioned, as it would constitute surveillance of the poor and race discrimination against Maori families which are subject to a disproportionate rate of child removals.[50]

[44] CARE 2012, p. 10 f.

[45] Ibid., p. 15.

[46] Ibid., p. 19 f.

[47] Ballantyne 2019, p. 20.

[48] CARE 2012, p. 32–34. The report recommended a full ethical evaluation, which was later conducted by Dare in 2013, see the report, p. 35 and Ballantyne 2019, p. 21.

[49] Ballantyne 2019, p. 20 f.

[50] Eubanks 2018, p. 138.

When a new minister of social development took office (from the same political party as her predecessor, the New Zealand National Party) the project was stopped in 2015.[51]

However, some of the researchers from the CARE team had won a contract to develop a similar PRM tool on the other side of the world in Allegheny County in Pennsylvania in the U.S.[52]

3.2 The Allegheny Family Screening Tool

The Children and Youth Service (CYS) in Allegheny County had been the source of public scandals, garnering national attention, which was in part related to a policy of preventing cross-racial adoptions, the Baby Byron case, and the homicide of toddler Shawntee Ford by her father, who had a record of violence and substance abuse that was known to the CYS.[53] Over the years, the CYS was also struggling with budget cuts.

The CYS had taken several measures to deal with the mounting problems. One of these measures was to create a data warehouse which would serve as a central repository, integrating information collected by the Department of Human Services, other county agencies and state public assistance programs. The data warehouse, eventually containing over more than a billion digital records, later proved useful as the foundation for designing and implementing decision support tools and predictive analytics. One idea was to build an automated triage system to help in setting priorities and making better use of the resources available to the CYS.[54]

The CARE team from New Zealand was assigned to design a PRM tool, similar to the Vulnerable Children PRM, using the data warehouse to harvest data in order to make predictions about probable maltreatment of children residing in Allegheny County.[55] The Allegheny Family Screening Tool (AFST) is linked to the county child abuse and neglect hotline, the ChildLine. Formerly the staff at the CYS were required to manually access and analyse vast amounts of data. This can now rapidly be performed by the AFST, which will produce a risk score regarding the long-term probability of future involvement in child welfare. The AFST

[51] Møller Jørgensen et al. 2021, p. 4 f.

[52] Eubanks 2018 p. 138.

[53] Ibid., p. 133.

[54] Ibid., p. 135–136.

[55] Ibid., p. 136–137.

is combined with other traditionally gathered information. If the score reaches a certain level, the CYS is obliged to initiate an investigation. According to the information on Allegheny County's website, the use of the AFST does not replace a clinical judgment but is used as additional information.[56]

The AFST has been a source of inspiration to other counties in the U.S. Nevertheless, it is far from uncontroversial. Concerns, similar to those faced by the Vulnerable Children PRM in New Zealand, have been raised regarding the AFST.[57] Even though the AFST shows the same degree of accuracy as its New Zealand counterpart, 76 percent in the area under the ROC, there is a great risk of harm to children and their families when a false positive occurs.

The use of proxies in the AFST, such as that of re-referrals (abuse notifications) is problematic, meaning that re-referrals are a variable no matter the reason for them. This is for example the case if several referrals are made regarding the same child either by someone with the aim to harass a parent or a family or due to so called "referral bias", which is often racially grounded.[58] According to various studies there is a disproportionately greater number of referrals concerning black or biracial families in Allegheny County.[59]

Similarly, there is also a class-based disproportionality concerning children placed in foster homes as a majority of placements concern families receiving different benefits for families in need. In conclusion, the use of public services appears to be considered a risk factor, in the same way as the Vulnerable Children PRM. In this regard, the tool is not designed to protect children from all class backgrounds against maltreatment. Furthermore, it has been criticized for being a tool for poverty profiling, confusing "parenting while poor with poor parenting".[60]

In Europe, similar PRM tools have been introduced by local governments in several countries.

[56] Information on the Allegheny County website: https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx.

[57] Allegheny County has rebutted the critique by Eubanks on their website, although without specifying any inaccuracies. See https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx.

[58] Eubanks 2018, p. 143, p. 153, p. 156.

[59] Ibid., p. 153.

[60] Ibid., p. 157–158.

3.3 The Hackney Early Help Profiling System

Hackney County Council in London, U.K., introduced an *Early Help Profiling System* (EHPS) in 2018 to help identify children at risk of neglect or abuse as part of the policy and practice of the Troubled Families Programme. The system is based on a predictive risk model bringing together data from multiple agencies. The underlying idea is that the children will be identified at an earlier stage before they come into contact with social workers, which will reduce costs.[61]

Scandals such as Baby P and Victoria Climbié, where small children already known to the authorities had been tortured and murdered by their caregivers, made it painfully evident that failures to share and act on information by the social services can have lethal outcomes. These scandals led to the idea of introducing so-called early help profiling systems in some municipalities in the U.K.[62] The scandals also contributed to new legislation, in the form of the Children Act 2004, which enhanced the possibilities of data sharing between agencies and provided local authorities with better access to information about the services that children in their respective areas were in contact with and contact information regarding the professionals involved. This was to be ensured by the application and synchronization of public databases.[63]

Part of this development was the online database RYOGENS (Reducing Youth Offending Generic National Solution) developed by the British Government together with consulting firm Deloitte and some other private companies. RYOGENS enabled officials from different agencies, such as Education, Police, Health Services, Social Services, Youth Offending Team and Housing Services to share information regarding a child at risk by filling in a form including forty different risk factors. If a certain threshold of reported concerns was reached, the system would generate an alert, which would be handled by a RYOGENS management function.[64]

[61] L. Dencik, A. Hintz, J. Redden & H. Warne, *Data Scores as Governance: Investigating uses of citizen scoring in public services*, Project Report, December 2018, Data Justice Lab, Cardiff University, U.K., p. 56.

[62] Dencik, et al. 2018, p. 58.

[63] R. van Brakel, *The Rise of Preemptive Surveillance: Unintended Social and Ethical Consequences*, Chapter 14 in E.R. Taylor and T. Rooney, Surveillance Futures: Social and Ethical Implications of New Technologies on Children and Young People, Routledge, London 2016, p. 189.

[64] Van Brakel 2016, p. 190.

The EHPS can be seen as yet another initiative "to explore the application of 'big data' solutions" regarding early intervention practises.[65] However, the EHPS was also built in the context of yet another harsh reality for the child services of Hackney Council, namely the combination of drastic funding cuts and an increase in the number of children on child protection plans and entering care.[66]

The EHPS was developed together with the private company Xantura, and funded by EY and London Councils.[67] The model integrated data from multiple agencies to identify children at risk of neglect or abuse in order to "strengthen the triage and assessment process"[68] and was expected to provide social workers with monthly risk profiles with integrated information about families with the greatest need of early intervention. The EHPS was therefore expressly said to be designed not to be punitive, only to enable earlier intervention.[69]

Only pseudonymized data was used by the model, meaning that data would only be made identifiable to the professionals assigned to deal with alerts generated by the model indicating that a high-risk threshold had been passed.[70]

No systematic account of the predictive variables in the model seems to be publicly accessible, but datasets that have been identified in a research study as well as in the media relate to school attendance, exclusion data, housing association repairs, arrears data, police records on anti-social behaviour and domestic violence, names, addresses, dates of births, unique pupil numbers, children and adult social care, housing debt, council tax, housing benefits and substance abuse data.[71]

[65] Ibid., p. 189–190.

[66] L. Stevenson, *Artificial Intelligence: how a council seeks to predict support needs for children and families*, Community Care, 1 March 2018, available at: https://www.communitycare.co.uk/2018/03/01/artificial-intelligence-council-seeks-predict-support-needs-children-families/.

[67] Dencik et al. 2018, p. 55.

[68] Information on the website of Xantura: https://xantura.com/early-help-profiling-system/.

[69] Dencik et al. 2018, p. 56.

[70] Ibid., p. 58.

[71] Ibid., p. 60; N. McIntyre and D. Pegg, *Councils use 377,000 people's data in efforts to predict child abuse*, The Guardian 16 September 2018, available at: https://www.theguardian.com/society/2018/sep/16/councils-use-377000-peoples-data-in-efforts-to-predict-child-abuse. Vannier Ducasse has expressed that information about the English

Regarding the accuracy of the EHPS, it was reported that over 80 percent of Hackney households identified as most at risk by the model were also at risk in real life.[72] According to media reports the EHPS helped detect seven children in need of early help support of whom Hackney Council was earlier unaware of.[73] A study presented in 2020 by What Works for Children's Social Care shows that there is no evidence that machine learning works satisfactorily in terms of accuracy when it comes to identifying children at risk.[74]

As a whole, the development procedure of the model lacked in transparency due to references to Xantura's commercial interests, which was presented as the reason to why several freedom of information requests (FOI) by researchers were denied.[75]

The fact that no information as to how many families were wrongfully identified as high risk and how those situations were handled, for example concerning the possibilities of removing such wrongful assessments from the EHPS, has been the subject of criticism.[76]

The entry into force of the GDPR as well as the Cambridge Analytica Scandal surely played a role in highlighting the data protection concerns voiced in the media regarding the EHPS, especially as the persons targeted by the EHPS were not informed of the use of their personal data and that no opt-out options were presented to them.[77]

The EHPS came to a halt in 2019 when it was concluded that the expected benefits would not be realized, which was mainly due to the lack of accuracy and data.[78] Looking forward a local politician, Darren Martin of the Hackney Liberal Democrats, stated:

experiments regarding PRM tools for child welfare and early intervention is "meagre", see Vannier Ducasse 2020, p. 4.

[72] Denick et al. 2018, p. 62.

[73] E. Sheridan, *Town Hall drops pilot programme profiling families without their knowledge*, Hackney Citizen, 30 October 2019.

[74] Møller Jørgensen et al. 2021, p. 5; Turner, A, '*No evidence' machine learning works well in children's social care, study finds*, Community Care, 2020 September 10, https://www.communitycare.co.uk/2020/09/10/evidence-machine-learning-works-well-childrens-social-care-study-finds/.

[75] Møller Jørgensen et al. 2021, p. 5; Dencik et al. 2018, p. 59–60.

[76] Møller Jørgensen et al. 2021, p. 6.

[77] Vannier Ducasse 2020, p. 19; Dencik et al. 2018, p. 62.

[78] Møller Jørgensen et al. 2021, p. 5; Sheridan 2019.

> …In a future where algorithmic technology will be used more and more, people have to know exactly how their data is being used… What we need now is an assurance that any future trial of this nature needs to be put in a public consultation with full disclosure of exactly what data is collected and how it will be used.[79]

3.4 The Gladsaxe model

The Gladsaxe model, based on a predictive algorithm to identify children at risk, received a great deal of national attention in Denmark.[80] It was established by a municipality in the suburbs of Copenhagen, inspired by prior models developed in New Zealand and USA, with the aim of creating an early warning system for detecting vulnerable children before they showed any symptoms of dysfunction.[81] A clear advantage of the model was that it could provide an overall assessment of the situation of the child through the mining of data from different sectors, with the potential of serving as a valuable supplement to professionals. If the model identified a child, a specialist adviser would make a preliminary assessment. If the expert found that there was reason to proceed, the family would be contacted and offered help. If the family declined, the municipality would not take any further action.[82]

The point-based model used data about several risk indicators such as mental illness (3000 points), unemployment (500 points), missing a doctor's appointment (1000 points) or dentist's appointment (300 points). Divorce was also included in the risk assessment. The model extracted data from nine different public sources, for example, the employment system used by job centres, the central personal register, dentist journals, the day care system and notifications of concern received by public authorities.[83]

[79] Sheridan 2019.

[80] Møller Jørgensen et al. 2021, p. 7.

[81] Ibid. p. 8. Møller Jørgensen et al. points out that the cross-national influence of the Vulnerable Children PRM is evident in Rhema Vaithianathan's inclusion in one of the scientific advisory boards of the project.

[82] U. Andreasson and T. Stende, *Nordic municipalities' work with artificial intelligence*, Nordic Council of Ministers 2019, p. 22, available at: https://www.norden.org/en/publication/nordic-municipalities-work-artificial-intelligence.

[83] R.F. Jørgensen, *Data and rights in the digital welfare state – the case of Denmark*, Information, Communication & Society 2021, p. 8, https://doi.org/10.1080/1369118X.2021.1934069.

The model was supposed to be rolled out in relation to all families with children within the municipality, but there were problems related to the accuracy of the model and a relatively high error rate, mainly due to the lack of historical data. The municipality had also made a request to the data protection agency to be exempt from the data protection legislation in order to access data from different sources, which was denied.[84]

The problems did not end there. When the model became public knowledge through an article in the daily newspaper Politiken, it caused a public outcry. The model was depicted as a tool for mass surveillance of families with children and the idea of a point-based system went above and beyond what was deemed to be acceptable.[85]

Nevertheless, the Danish Government was ready to go through with a legislative proposal which would allow municipalities in Denmark to combine data regarding families with children and children in general as part of an overarching plan to fight parallel societies, also known as the "ghetto-plan", which would enable the scoring of neighbourhoods. If a neighbourhood scored high enough to be qualified as a ghetto, several measures would be put into place, such as applying automated risk assessment systems to families with children. The proposal was later withdrawn.[86]

The Gladsaxe model, as its focus was not only in relation to a part of the population receiving benefits but to the entire population, can be said to be an example of a model that generally had a broader reach than the Vulnerable Children PRM or the Allegheny Family Screening Tool.

3.5 The Norrtälje model

In 2020, the municipality of Norrtälje became the first in Sweden to develop a tool using a Robotic Process Automation system (RPA) involving AI to identify children at risk. The system would collect and analyse previous cases as a tool to help social workers make a decision concerning the initiation of a child protection investigation after receiving reports

[84] Ibid.

[85] B. Alfter, *Denmark* in *Automating Society – Taking Stock of Automated Decision-Making in the EU* – A report by AlgorithmWatch in cooperation with Bertelsmann Stiftung, supported by the Open Society Foundations, 1st edition, January 2019 p. 51; see also J. Sorgenfri Kjaer, https://politiken.dk/indland/art6365403/Regeringen-vil-overvåge-alle-landets-børnefamilier-og-uddele-point.

[86] Alfter 2019, p. 51; Andreasson and Stende 2019, p. 22.

of concern (*orosanmälningar*). The project was essentially funded by the municipality with some help from Vinnova, Sweden's innovation agency. Part of the background concerning the pilot project was the 50 percent increase in reports of concern between 2014 and 2018. There was an urgent need for support measures for the social services.[87]

The system basically works through a web-service for concern reports, where digital reports will be received from the mandatory reporters, such as schools, health care and police, in a structured manner. Next, the AI tool will read and analyse the information received via the web-service as well as registering it in the operating system. Finally, it will create a pre-assessment proposal (*förhandsbedömning*) via a predictive model tool for pattern recognition based on prior assessments. A child welfare officer will decide whether the pre-assessment proposal will be documented.[88]

The dataset includes anonymized administrative data related to all prior assessments regarding the initiation of a child protection investigation by the municipality. The model is designed to compare words in new reports with earlier reports and to make an assessment based on the latter reports.[89] Björn Preuss from the company *2021.AI* has provided the following explanation:

> We do not select any information manually or include any factors. We only use the text which is sent with every report. The only information which is prior to the model detected and filtered away is personal information like names, age, social number, etc. So the model cannot be biased towards a name, gender, age etc. All predictions are only based on historic text descriptions of cases and their statistical similarity, word and sentence patterns, etc.[90]

[87] See official statement by the IT-department at Norrtälje municipality regarding IT-investment, a platform for automation and decision support, 2019-07-17, available at: https://forum.norrtalje.se/welcome-sv/namnder-styrelser/kommunstyrelsens-arbetsutskott/mote-2019-08-28/agenda/tjansteutlatande-gallande-investering-for-plattform-for-automatisering-och-beslutsstodpdf-35012?downloadMode=open; *Larmet: 200 barn om dagen misstänks fara illa i Stockholms län*, SVT 17 February 2020, https://www.svt.se/nyheter/lokalt/stockholm/200-barn-om-dagen-orosanmaldes-under-2018.

[88] *Projekt för AI och robotisering av orosanmälan*, information on the Norrtälje municipality website, available at: https://www.norrtalje.se/ai_oro; P. Molander Wistam, PowerPoint Presentation 23 August 2021, RPA/AI Flödesbeskrivning.

[89] F. Adolfsson, *AI för Norrtäljes orosanmälan*, Voister 13 november 2019, available at: https://www.voister.se/artikel/2019/11/ai-for-norrtaljes-orosanmalan/.

[90] Quote from an e-mail from P. Molander Wistam, 24 August 2021. Also see M. With, Dansk IT 12 October 2020, *AI for the sake of the children*; the client case of 2021.AI:

After a legal review supported by the Swedish Association of Local Authorities and Regions, the municipality decided not to pursue the project.[91] Swedish law as such does not prevent the use of predictive modelling regarding historical data about individuals for the purposes of case management and for developing quality assurance within the social services.[92] There are, however, important limitations as to how this can be carried out. Search limitations according to the law, include, for example, automated data processing regarding reports of concern or pre-assessments which did not lead to a decision to initiate a child protection investigation, even if the child already had a case file at the social services.[93] This posed a problem regarding the deployment of the Norrtälje model, since it was necessary to use such pre-assessments as a source. The legal limits at hand have been the subject of debate for decades but the issue has repeatedly been dismissed as contrary to the right to the protection of privacy. However, the issue is not off the table and new legislative proposals are being considered by government authorities.[94]

The Norrtälje model does not seem to have been subjected to any research analysis thus far, but important research regarding automation in social services lay bare some of the important challenges that the use of historical administrative data, such as prior decisions, within an AI tool might entail. This concerns the possible cementing of former biases as well as the balancing of interests in individual cases, which is required by the principles of the rule of law.[95]

Applying AI to create more comprehensive, safe and accurate assessments of social service cases in Norrtälje Municipality available at: https://dit.dk/nyheder/2020/for-the-sake-of-the-children.

[91] A. Yanchur, G. Rosén Fondahn and S. Pilz, *A Swedish town bought an AI to spot children at risk, but decided against deploying it*, Algorithm Watch 10 August 2021.

[92] M. Nymark, *Användning av AI inom socialtjänsten*, report, Swedish Association of Local Authorities and Regions 2021-02-07, available at: https://skr.se/download/18.427140af179361c4e4616b7a/1620377226836/Anv_%20av_%20AI_%20inom_%20socialtjansten_%20rapport.pdf.

[93] Nymark 2021, p. 18.

[94] Socialstyrelsen (the National Board of Health and Welfare), *Att göra anmälningar som gäller barn sökbara*, Report May 2019, available at: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/ovrigt/2019-5-15.pdf.

[95] L. Svensson, *Automatisering – till nytta eller fördärv?* (Automation – benefit or harm?) Socialvetenskaplig tidskrift 2019:3-4, p. 358–359.

3.6 Concluding remarks

This preliminary overview of examples of the use of AI tools for child protection in social services reveals that most of them have been created in the context of struggles related to increases in caseloads, funding cuts and staff shortages in relation to social services as well as government ambitions to increase digitalisation in the public sector. Thus, the primary reasons for developing such tools mainly seem to be of a financial and administrative nature.

This overview also shows that many of the projects developing AI tools for child protection have been discontinued at the experimental stage, which mainly seems related to legal, ethical and public trust problems. Legal limits as well as state-structures can limit the amount of data that can be used in a model, which can render it more or less useless. The only tool that has survived so far is the Allegheny Family Screening Tool in the U.S.

The tools also differ in purpose and scope. Some of them use point-based systems related to individuals regarding certain characteristics or activities, while others use text-mining.

Nevertheless, this is an ongoing trend which is presumably here to stay.

4 Children's rights, child protection services and AI tools

4.1 Introduction to the children's rights system in Europe

AI-tools for child protection can have huge legal implications, in particular concerning children's rights, and have the power to severely impact the lives of children. In the end, however, it all comes down to how AI-tools are used and for what purposes. A framework for the use of AI-tools, at a minimum, needs to be developed that is in accord with children's rights. The child is an independent rights holder.[96] The focus of this paper is a European Human Rights perspective.

One of the most important children's rights instruments is the United Nations Convention of the Rights of the Child (UNCRC) adopted in

[96] W. Vandenhole, G. Erdem Türkelli and S. Lembrechts, *Children's Rights: A Commentary on the Convention on the Rights of the Child and Its Protocols*, Edward Elgar Publishing Ltd 2019, p. 15.

1989. It has been ratified by all the members of the UN except the U.S. It is binding for the signatory states. In states with monist systems, the UNCRC is directly applicable, whereas in states with dualist systems the applicability depends on whether that state has incorporated the UNCRC as a part of national law,[97] which is for example the case in Sweden since 2020.[98] The fact that the UNCRC is binding on the signatory states does not mean that it is enforceable by individuals. There is no international court of children's rights to turn to, and no other court unless a signatory state has decided to make the rights enforceable in a national court of law. Nevertheless, the Committee on the Rights of the Child has both a monitoring and advisory function.

More importantly the UNCRC is highly integrated into the European human rights system. The member states of both the Council of Europe and the EU are parties to the UNCRC, and the UNCRC has been described as "the touchstone for the development of European children's rights law".[99]

This development has mainly taken place within the European Convention on Human Rights (ECHR) framework. The ECHR from 1950 applies in most states in Europe, is a part of EU law, and provides an enforceable protection of children's rights through the European Court of Human Rights (ECtHR). The case law of the ECtHR has had an important practical impact on children's rights in Europe, including numerous cases regarding child protection,[100] even though the application of the principle of the margin of appreciation for the states has been a focus of criticism in cases related to "the best interest of the child".[101]

Inspired by the UNCRC,[102] children's rights are also regulated in Art. 24 of the EU Charter of Fundamental Rights (EUCFR) of 2009, but the scope is limited to certain cross-border situations, such as criminal law

[97] Vandenhole et al. 2019, p. 21.

[98] Prop. 2017/18:186; See also K. Åhman, P. Leviner, K. Zillén (ed.) *Barnkonventionen i praktiken – Rättsliga utmaningar och möjligheter*, Norstedts Juridik Poland 2020 p. 30–42 and Grahn Farley 2019 p. 26–28.

[99] *Handbook on European law relating to the rights of the child*, European Union Agency for Fundamental Rights and Council of Europe 2015, p. 26.

[100] Vandenhole et al. 2019, p. 18.

[101] R. Lamont, *Article 24 – The Rights of the Child* in S. Peers, T. Hervey, J. Kenner and A. Ward (eds), The EU Charter of Fundamental Rights – A Commentary, Hart Publishing 2014, p. 673.

[102] Lamont 2014, p. 674.

and immigration law. The reason for this is that the EU does not have any direct general competence regarding children's rights.[103] It will therefore not be further examined in this paper. Nonetheless it is noteworthy that the EU commission is actively working with children's rights and introduced a strategy on the rights of the child and the European Child in March of 2021. The strategy developed has been guided by the UNCRC with the purpose of securing access to basic services for vulnerable children. An important aim of the strategy is to break vicious cycles across generations related to child poverty and social exclusion.[104]

The analysis below will focus on the rights of most relevance to the use of AI-tools for child protection. This includes rights with a direct purpose of protecting children from maltreatment, the right to life and the prohibition of inhuman and degrading treatment, as well as the right to respect for family life in conjunction with the prohibition of discrimination. These tools have the capacity to both enhance and/or interfere with children's rights, which will be elaborated below.

4.2 The child's right not to be maltreated and the positive obligation for the state to make risk assessments

It can be said that the utmost duty to protect children rests upon the state. If parents or other legal caregivers are not able or unfit to take care of children, *i.e.* human beings under the age of 18 as prescribed in Art. 1 of the UNCRC, the state is required to intervene. In Art. 3.3. of the UNCRC this is expressed as:

> State Parties shall ensure that the institutions, services and facilities responsible for the care or protection of children shall conform with the standards established by competent authorities, particularly in the areas of safety, health, in the number and suitability of their staff, as well as competent supervision.

There is thus a *positive obligation* for the state to protect children from maltreatment, i.e., circumstances when a State has a duty to take action

[103] Lamont 2014, p. 662.

[104] Communication from the Commission to the European Parliament, the Council and the European Economic and Social and the Committee of the Regions, *EU strategy on the rights of the child*, Brussels 23.4.2021 COM (2021) 142 final.

in order to secure the protection of individuals within its jurisdiction,[105] which can involve complex risk assessments. When this turning point is reached is a delicate matter requiring a complicated balancing act which involves the human rights of both the child and the caregivers.[106] The idea that AI-based tools for risk assessments could be used to help in making such risk assessments therefore seems highly relevant.

If the state fails to protect a child a whole range of human rights come into play of both an absolute and relative nature, raising the question, to what extent are there exceptions to a right. In extreme cases such as Baby P and Victoria Climbié in the U.K. and the case of "Little heart" (*Lilla hjärtat*) in Sweden, where small children already known to the authorities, have died at the hands of their caregivers, the right to life laid down in Art. 2 of the ECHR and Art. 6.1 UNCRC, which is an absolute right, is applicable if the state did not act on the evidence or information available to them.

The ECtHR applies the so-called Osman-test to assess whether state authorities have taken the necessary preventive measures in cases where children are at high risk, i.e. when the positive obligation of the state is triggered:

> It must be established…that the authorities knew or ought to have known at the time of the existence of a real and immediate risk to the life of an identified individual or individuals from the criminal acts of a third party and they failed to take measures within the scope of their powers which, judged reasonably, might have been expected to avoid that risk.[107]

This threshold can be met in the case of domestic abuse against a parent who is known to the authorities, since this means that the child is at a high risk of maltreatment. For example, in the cases of *Kontrova v. Slovakia*[108] and *Talpis v. Italy*[109] where women had reported serious abuse and threats with lethal weapons by their partners to the police, the failure of the police to investigate and report to the social services led to the killing

[105] B. Rainey, E. Wicks, and C. Ovey, *Jacobs, White & Ovey – The European Convention on Human Rights*, 6th ed., Oxford University Press, U.K. 2014, p. 103.

[106] See e.g. *Z. and Others v. the United Kingdom*, n° 29392/95, Judgment (GC) 10 May 2001, § 74.

[107] *Osman v. The United Kingdom*, n° 23452/94, judgment (GC) 28 October 1998, § 115.

[108] *Kontrova v. Slovakia*, n° 7510/04, Judgment 31 May 2007.

[109] *Talpis v. Italy*, n° 41237/14, Judgment 2 March 2017.

of minor children in the family. The ECtHR found that the state had violated the right to life regarding the children. In the Talpis case the court concluded that:

> Article 2 of the Convention may also imply in certain well-defined circumstances a positive obligation on the authorities to take preventive operational measures to protect an individual whose life is at risk from the criminal acts of another individual.[110]

The ECtHR did not find that the authorities had made a correct risk assessment in the Talpis case. Adding to the Osman-test the Court stated that:

> In the Court's view, the risk of real and immediate threat must be assessed taking due account of the particular context of domestic violence. In such a situation it is not only a question of an obligation to afford general protection to society…but above all to take account of recurrence of successive episodes of violence within the family unit.[111]

The ECtHR found that the state had failed to live up to its positive obligations to take preventive operational measures to protect an individual whose life is at risk.

In some situations, however, it can be difficult or even impossible to foresee the killing of a child by his or her caregiver, such as in the case of *Penati v. Italy* where a father had killed his son and himself during a protected contact session between the father and son on the premises of the social services of a municipality. As long as the authorities have taken the necessary preventive measures that are available, they cannot be held liable for a violation of the right to life.[112]

The Grand Chamber judgement in *Kurt v. Austria*, where, following an escalating spiral of domestic violence involving both the mother and the children, the father shot his 8-year-old son to death at school, provides further clarifications regarding the Osman-test in the form of general principles. In this case, however, the dissenting opinion shows that the judges were not in agreement with each other on where to draw the line as to what can be demanded of the authorities when it comes to risk

[110] *Talpis v. Italy*, § 101.

[111] *Talpis v. Italy*, § 122.

[112] *Penati v. Italy*, n° 44166/15, Judgment 11 May 2021, § 188 (available only in French). The applicant has requested a referral to the Grand Chamber.

assessments regarding domestic abuse cases, and in particular the risk for lethal outcomes.

The court established that when there is a real and immediate risk to the life of a victim of domestic violence, the authorities have a duty to carry out a lethality risk assessment in an autonomous, proactive and comprehensive manner. Nevertheless, the Osman-test does not require that states use standardised risk assessments, such as standardised checklists based on criminological research, even though the court acknowledged that such assessments are useful.[113] The court also concluded that in the case where "several persons are affected by domestic violence, be it directly or indirectly, any risk assessment must be apt to systematically identify and address all the potential victims."[114] It also emphasised the importance of documentation, information sharing and coordinated support with other relevant stakeholders that come into regular contact with persons at risk, which in the case of children can be teachers. The authorities should also communicate the outcome of their risk assessment to the victims and, when necessary, give advice and guidance regarding different protective measures available to them.[115]

By ten votes to seven, the majority held that Austria had met these requirements in the Kurt-case and that there thus had been no violation to the right to life in this case.

The minority, however, found that the risk assessment was seriously flawed and that the State had breached the right to life. Among others, the minority pointed out that the authorities failed to make a separate risk assessment in relation to the children and did not treat the risk of domestic violence as one that impacted the family as a unit. This was particularly grave since the authorities had information which indicated a high risk to the children. Apart from statements given by the children themselves regarding physical abuse by the father, the authorities evidently downplayed the fact that the father had made explicit and repeated threats to the mother that he would kill the children.[116] The lack of standardized research-based assessment tools by the authorities was highlighted in this regard.

[113] *Kurt v. Austria*, n° 62903/15, Judgment (GC) 15 June 2021, §§ 168-171.

[114] *Kurt v. Austria*, § 173.

[115] *Kurt v. Austria*, § 174.

[116] *Kurt v. Austria*, Joint dissenting opinion by judges Turkovic, Lemmens, Harutynyan, Elósegui, Felici, Pavli and Yüksel, § 13.

A more common scenario is that if there is enough evidence to support that a child has been subject to torture, abuse or neglect, the authorities are obliged to act and thoroughly investigate such a case and, if necessary, take the appropriate measures. A failure to act, can constitute a breach of Art. 3 ECHR which includes the prohibition of torture or other inhuman or degrading treatment. Art. 19.1 of the UNCRC stipulates that:

> State parties shall take all appropriate legislative, administrative, social and educational measures to protect the child from all forms of physical or mental violence, injury or abuse, neglect or negligent treatment, maltreatment or exploitation, including sexual abuse, while in the care of parent(s), legal guardian (s) or any other person who has care of the child.

A crucial factor in this regard is the degree of maltreatment. The court has found that for maltreatment to fall within the scope of Art. 3 ECHR, the maltreatment must attain a minimum level of severity. To this end an overall assessment of the relevant circumstances of the case has to be conducted, taking into consideration, for example, the nature and context of the treatment, its duration, its physical and mental effects, and in some cases the sex, age and state of health of the victim.[117]

In cases regarding neglect or abuse, the need for authorities to act swiftly is crucial. The ECtHR has in numerous judgments criticised states for the failure to act on information available to them. In the case of *Z. and Others v. U.K.* repeated concerns had been reported to the social services about a family with four small children during a period of four and a half years. The children had been subjected to severe neglect and emotional abuse, where the parents kept the children locked up in their rooms which were extremely filthy, or locked them out of the home. The children were malnourished, dirty and were regularly caught stealing food from bins. It was not until the mother demanded that the social services put the children up for adoption and care, as she could not cope with them, that the children were taken in for emergency care. The Court found that, in the present case, it was not in dispute that the neglect and abuse suffered by the children reached the threshold of inhuman and degrading treatment. It was concluded that the authorities were under a statutory duty to protect the children and had a range of powers available to them, which included the removal of the children from their home. The Court acknowledged that the social services are faced with a diffi-

[117] *Costello-Roberts v. the United Kingdom*, n° 13134/87, Judgment 25 March 1993, § 30.

cult and sensitive task to balance the duty to uphold the countervailing principle of respecting and preserving family life and assessing the risk of maltreatment. Nevertheless, in the present case, there was no doubt as to the failure of the system to protect the children from serious, long-term neglect and abuse.[118]

In the case of *E. and others v. the U.K.* three sisters and a brother had been subjected to long-term, severe, physical and sexual abuse by their mother's partner. The partner was convicted of sexually assaulting two of the girls. When he came back to live with the family while on probation, the authorities failed to take the necessary steps to monitor and supervise the family and make the necessary risk-assessments, which meant that the abuse could continue for several years. The children suffered serious mental disorders as a result. The Court made the assessment that the state had not reasonably used the measures available. There was a clear pattern of a lack of investigation, communication and co-operation by the relevant authorities which would have had the possibility to avoid or at least minimize the risk of the damage suffered.[119]

A specific situation where the positive obligation of the state is normally triggered is, for example, when a head teacher reports concern of suspected maltreatment. This is especially the case since such a report presumably is reflective of teachers who have the child or children concerned on their watch on a daily basis. The authorities are hereby obliged to take the necessary precautionary measures, including a child maltreatment risk assessment.[120]

In sum, the case-law of the ECtHR provides us with general principles regarding the maltreatment risk assessment and lethality risk assessment. Suspicions of maltreatment and or risk for the child's life will trigger the immediate need for appropriate measures to be taken. The duty to trace child maltreatment is somewhat vague, but Art. 3 ECHR and Art. 19 UNCRC require legislative and administrative measures, as well as social and educational measures to be in place. Certainly, institutions such as schools and school health services play an important role in the detection

[118] *Z. and Others v. the United Kingdom*, n° 29392/95, Judgment (GC) 10 May 2001, § 74.

[119] *E. and Others v. the United Kingdom*, n° 33218/96, Judgment 26 November 2002, §§ 99-100.

[120] *Association Innocence en Danger v. France and Association Enfance et Partage v. France*, n° 15343/15, 16806/15, Judgment 4 June 2020 (available in French and German), § 161, § 167.

of child maltreatment. The right of the child not to be maltreated however does not at this point in time seem to encompass the prediction of child maltreatment in cases where there is no "smoking gun". States are however encouraged to use research-based, multidisciplinary risk assessment standards for the prevention and mitigation of child maltreatment.

Consequently, the use of AI-tools for child protection may have the potential to make the risk assessment process by the relevant authorities more effective, which in turn may enhance the protection of children from maltreatment and prevent death. Nevertheless, this requires that the system is legal, research-based and has a high degree of accuracy. In the light of Art. 22 GDPR it is also important that AI-driven child protection tools will be used in such a way that the experts will not solely rely on such tools. In a study conducted by Bosk, it was determined that one third of the social workers were positive to using a risk score, in part because it was seen as an important tool to prevent subjective decision making and perhaps more noteworthy in part because it "removed the responsibility (and terror) of making a mistake". If social workers would start to rely solely on risk scores, this could in practise constitute illegal automatic decision-making. Instead, they could serve as part of an elaborate method using several different tools. Moreover, it is important that such an AI-tool is developed in a proper manner including the examination of various risk factors, which means that issues regarding discrimination in particular must be assessed.

4.3 The child, the right to respect for family life and the prohibition against discrimination

If social services decide to take measures that can be more or less intrusive into the family life of the individuals involved or even separation of the family members, the right to family life stipulated in Art. 8 of the ECHR has to be considered.[121] This involves both the child and the caregivers, such as biological parents or foster parents.[122]

A primary consideration in this regard is the somewhat vague concept of "the best interest of the child" Art. 3.1 UNCRC, which is applied by

[121] *Strand Lobben and Others v. Norway*, n° 37283/13, judgment (GC), 10 September 2019, §§ 202-04.

[122] See e.g. *Kopf & Liberda v. Austria*, n° 1598/06, Judgment 17 January 2012 regarding the right of respect to private and family life of foster parents (Art. 8 ECHR).

the European Court of Human Rights as well as in many national legal systems. In this context it has the power to override the rights of the parents, since the aim pursued regarding child protection measures is the best interest of the child.[123]

As stated above, decisions regarding measures such as early intervention are a delicate matter. They involve issues such as what constitutes good or adequate parenting, which might give rise to discriminatory assessments based on factors such as socioeconomic status, the level of education of the parents, disabilities or illnesses, place of residence, race, religion, culture etc. Social services are thus required to work to prevent bias from being part of the decision-making process, which might prove particularly difficult when using AI driven tools. This is a hurdle that has to be overcome in an effective manner if such technology is to be used in the first place. A general prohibition of discrimination is regulated in Art. 14 of the ECHR and more specifically in Art. 2.1 of the UNCRC, which reads:

> State parties shall respect and ensure the rights set forth in the present Convention to each child within their jurisdiction without discrimination of any kind, irrespective of the child's or his or her parent's or legal guardian's race, colour, sex, language, religion, political or other opinion, national, ethnic or social origin, property, disability, birth or other status.

This article leaves room for a broad interpretation of what constitutes discriminatory treatment by the state. Article 14 ECHR is not applied independently but will be applied in conjunction with another right stipulated in the ECHR and is thus regarded as an ancillary right.[124] In the context of child welfare measures Art. 14 is often applied together with the respect to private and family life laid down in Art. 8 ECHR. Moreover, the protection against discrimination in Art. 14 is completed by Article 1 of Protocol No. 12 to the ECHR, which prohibits discrimination more generally, in the enjoyment of any right set forth by law. It is noteworthy that only 20 states among the signatory states have ratified Protocol No. 12.

[123] See e.g. *Vojnity v. Hungary*, n° 29617/07, Judgment 12 February 2013, § 43.

[124] *Guide on Article 14 of the European Convention on Human Rights and on Article 1 of Protocol No. 12 to the Convention – Prohibition on Discrimination*, Updated on 31 December 2020, Council of Europe/European Court of Human Rights, p. 6.

The risk of discriminatory assessments in this context is mainly related to the parents, which can include both characteristics and behaviour. To use characteristics as risk variables is therefore especially risky in regard to the prohibition against discrimination. It can also raise issues regarding so-called intersectionality, that is, the interplay of several grounds of discrimination at the same time, such as social background, sex, race, ethnicity, sexual orientation, disability, and age. In such cases there is a need for a more holistic and flexible approach, which cannot be satisfied by the use of single comparators.[125]

Several cases in the ECtHR case law are illustrative of this. The Court has criticised decisions to remove children from their parents solely on reasons of poor housing and poverty as contrary to the right to respect for family life. In some of these cases the measures notably targeted families where the parents had a certain ethnic background or disability.[126]

It has also been deemed contrary to Art. 8 and Art. 14 ECHR to base the withdrawal of parental rights or parental access rights solely on the ground of disability,[127] mental illness,[128] religious considerations[129] or sexual orientation of the caregivers.[130] The case law, however, does not indicate that factors such as social background, disabilities or religious conviction of the parents cannot be part of an overall assessment of the parent-child relationship in cases where there are other circumstances such as a risk of abuse and neglect.

The question is what predictive risk variables would be lawful or appropriate to use when developing an AI-tool that will be constructed on the basis of many risk variables which have the potential to provide an important overview of a child protection case. This also raises issues

[125] S. Atrey, *Comparison in intersectional discrimination*, Legal Studies, 2018, 38, p. 379–395.

[126] *Barnea and Caldararu v. Italy*, n°, Judgment 22 June 2017 (Roma origin); *Saviny v. Ukraine*, n° 39948/06, Judgment 18 December 2008 (blind parents); *Wallová and Walla v. Czech* Republic, n° 23848/04, Judgment 26 October 2006, §§ 71-72.

[127] *Kocherov and Sergeyeva v. Russia*, n°16899/13, Judgment 29 March 2016 and *Kutzner v. Germany*, n° 46544/99, Judgment 26 February 2002 (mental disabilities).

[128] *Cînta v. Romania*, n° 3891/19, Judgment 18 February 2020, §§ 47-57 (paranoid schizophrenia).

[129] *Vojnity v. Hungary*, application n° 29617/07, Judgment 12 February 2013 and *Hoffmann v. Austria*, judgment 23 June 1993 (Parents belonging to a Pentecostal Charismatic Church and Jehovah's Witnesses respectively).

[130] *Salgueiro da Silva Mouta v. Portugal*, judgment 21 December 1999 (homosexual parent).

regarding intersectionality, which in the context of AI-tools might prove to be a major hurdle, since AI-tools can only make automatic pre-assessments based on certain risk factors.

The use of predictive variables is especially problematic when it comes to measuring the predictive risk variables in relation to outcome variables. It does not seem that there is any data mining process, such as the statistical procedure referred to as *stepwise probit or logistic regression* (SPLR) used in the Vulnerable Children PMR, that is certain to produce meaningful correlations.[131]

Instead, there is a clear risk that such correlations may be exaggerated or irrelevant. SPLR for example does not take into account the distribution of factors related to maltreatment in the rest of the population. The fact that parents have learning disabilities or health problems does not in itself mean that they cannot provide good parenting. If 10 percent of this group of parents maltreat their children, there is still 90 percent that does not. The weight given to such factors therefore poses a problem. If five percent in the general population would be considered as maltreating their children, the weight given to learning disabilities or poor health would double the child's risk of maltreatment. However, in absolute terms the risk is much lower regarding children with parents facing such problems.[132]

Furthermore, an SPLR method may create misleading results, since "any factor that varies with maltreatment is taken to be theoretically suitable and to enhance" the PRM. It fails to assess the degree of these factors, as they do not occur only in abusive families. The use of such methods can therefore not be considered to encompass the complexity of a balanced assessment regarding child maltreatment[133] and has for this reason been labelled a "statistical fishing expedition".[134]

Considering the Vulnerable Children PRM and the Allegheny Family Screening tool, it is clear that there is a direct connection to the child's or the caregivers' social origins and property or lack of property. Indirectly there are issues related to, for example, race and/or ethnic origin, since these grounds are often linked to the fact that due to prior discrimina-

[131] Eubanks 2018, p. 144.

[132] Vannier Ducasse 2020, p. 7.

[133] Ibid.

[134] Eubanks 2018, p. 144.

tion, certain groups in society have been oppressed and as a result of that also belong to less advantaged socioeconomic groups.

AI-tools of this kind that only include children whose parents are on welfare benefits seem unlikely to be in accordance with the prohibition against discrimination.

In the light of the prohibition against discrimination, the use of several other risk factors are problematic concerning a PRM tool, regarding both the child and the caregivers. Even though, for example, a religious conviction might pose a risk for child maltreatment if the parents adhere to a religious sect,[135] it is hard to see how this would be handled within an AI-tool, with all the different dimensions that might have to be assessed.

In conclusion, the use of PRM-tools to prevent child maltreatment do not seem suitable for making decisions regarding pre-assessment in child protection cases. These are decisions which require empathy, flexibility and intuition.

The NLP/TP model developed by Norrtälje Municipality, however, does not seem to be as problematic as the PRM-models in relation to direct discrimination. However, there is a risk that the status quo bias in former decisions will be included, which may lead to the repetition of biased or unrepresentative decision-making amounting to unlawful discrimination. Moreover, having in mind the evolution regarding both research and values regarding the child-parent relationship of the past two decades, European perceptions of family have undergone important changes, not least regarding lesbian, gay, bi- and transgender families as well as the role of fathers in children's development. It is clear that the area of child-parent relations is a dynamic area, which will undoubtedly lead to different assessments regarding the best interest of the child and not least concerning child maltreatment assessments in the light of the principle of evolutive interpretation of the ECtHR.[136] This has to be accounted for when developing and using a predictive tool based on AI.

[135] See e.g. *Tlapak and Others v. Germany*, application n° 11308/16 and 11344/16, Judgment 22 March 2018 (practices of caning within a religious sect).
[136] Rainey, Wicks, and Ovey, 2014, p. 73–78.

4.4 A preliminary outline of the legal issues related to AI tools for child protection

The design and use of AI-tools for child protection raises several legal issues that can be identified from the discussion of the AI driven tools that relate to the rights of the child. They pose problems that need to be overcome or dealt with.

To start, it is important to note that the rights of caregivers can also both directly and indirectly affect the child, which is why it is not entirely possible to apply a child centred approach without involving the family to some degree.[137] There can also be a question of maltreatment outside of the family by other adults or other children in the child's vicinity.[138]

The AI tools can result in the profiling of families with children based on for example racial, socioeconomic- and health status, which directly or indirectly targets the child. It has been shown that statistical methods can lead to wrongful outcomes since the correlations that they produce can be both exaggerated and irrelevant. Furthermore, a tool can be constructed for screening of large parts of a population, such as families with children, which can amount to mass surveillance that can be invasive not only for the parents but also affect the child in a negative way. This may be contrary to the right to respect for privacy and family life as well as protection of personal data and can undermine public trust, with the effect that parents as well as children may avoid seeking help from the authorities when in need.

The AI tools will likely include the biases related to their developers, which often can be related to race, gender, culture and socioeconomic status. This can especially be the case if the tool is designed to target only the part of the population that receives welfare benefits. It also risks cementing such biases into future decision-making. Consequently, there is a risk that such outcomes will amount to unlawful discrimination. Furthermore, such tools include the risk of excluding children at high risk who can be found in other socioeconomic groups in society.

Concerns have been raised relating to the opacity of the AI tools, the "black box problem", which can cause difficulties in understanding the reasons for an outcome that might serve as the basis for decision-making. In this context there is a conflict with the right to a fair trial in Art. 6

[137] Van Bueren 2018, p. 86.

[138] Ibid., p. 84.

ECHR, including the right to a motivated decision. More importantly, how do you contest such a decision legally and who will make sure that the child will be represented and by whom?

It is also important to note that the automated decision-making can be problematic if, in reality, there is no meaningful human involvement and oversight. If there is a risk that the staff rely too heavily on the outcome of an AI-tool and do not undertake any further controls, there is a risk that the child is subjected to an automated decision-making against the law and particularly Art. 22 GDPR.[139]

AI tools are never a hundred percent accurate when it comes to identifying child maltreatment, and the law requires a certain level of proof particularly regarding child protection measures that are by definition an interference in the right to privacy and family life.

Transparency problems might also arise if a state or local authority develops models together with private, for-profit entities. Access to information regarding the tools can be at risk, due to commercial interests and intellectual property rights, which in turn might be necessary information in a court of law if a decision based on the tool is contested. Moreover, the lack of transparency poses problems concerning trust and a sense of fairness for the child, youth and her or his care givers, which might lead the child to turn against society.

Last but not least, who will be accountable and held liable when an AI tool fails? And what reparations in regard to the child can be expected?

5 Final remarks with a view to the future

Tools using AI to counter child maltreatment may have the potential to enhance risk-assessments and serve as valuable decision-making support regarding child maltreatment. This certainly needs to be further researched. There are also other issues such as how the tools are supposed to be used, what procedures are elaborated in relation to the use of such tools, who will be qualified to make assessments using such tools and how will evaluations be carried out etc.? This certainly requires comprehensive regulation.

[139] Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and profiling for the purposes of Regulation 2016/679*, 3 October 2017 as last revised and adopted on 6 February 2018, WP251rev.01, p. 21.

As has been shown in this paper, there are several legal concerns that have to be addressed before designing, developing and using AI-tools to detect and prevent child maltreatment. It can therefore be concluded that there is a need to develop a children's rights framework for the use of artificial intelligence for child protection, a framework that can be included in a broader strategy regarding sustainable use of AI.

Furthermore, government AI-tools for the prevention of child maltreatment will need to be "future-proofed". The European Commission introduced a proposal for an Artificial Intelligence Act (AI-Act) in April 2021.[140] At present, it is not clear if or when the AI-Act will be adopted, but it is most likely that some kind of regulation will be adopted in a not-too-distant future. This will set further limits on the use of AI-tools concerning public child protection measures, especially regarding data quality and procedures for risk management. Considering Art. 5 of the AI-Act, most of the tools analysed in this paper, would probably be at risk of being prohibited since they involve social scoring (recital 17) and/or will probably be defined as "high-risk", due to the risk of harm particularly in relation to the fundamental rights of individuals.

[140] Proposal for a Regulation of the European Parliament and of the Council, Laying down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Brussels 21.4.2021, COM (2021)206, 2021/0106 (COD).

Stefan Larsson & Jonas Ledendal

AI i offentlig sektor: Från etiska riktlinjer till lagstiftning

1 Inledning

Användningen av artificiell intelligens (AI), det vill säga en rad primärt databeroende metoder och teknologier för bland annat prediktion och automation, tycks bidra till en omvälvande period i den offentliga förvaltningen. Detta skifte föranleder ett styrningsbehov, ofta kopplat till tillitsfrågor, vilket i en internationell kontext inte minst syns i en stor mängd etiska riktlinjer och principiella dokument som publicerats under de senaste få åren. Styrningen befinner sig därmed i en formativ period, vilket inte minst tydliggörs i och med EU-kommissionens förslag på den förordning om harmoniserade regler för AI (rättsakt om AI) som publicerades i april 2021.[1] Vi vill här belysa denna styrningsutveckling i sig, visa på dess mest centrala delar, samt analysera den svenska offentliga AI-användningen i dess ljus.

1.1 Regeringsuppdrag i ljuset av internationell AI-styrning

I juni 2021 fick Myndigheten för digital förvaltning (DIGG), Arbetsförmedlingen, Bolagsverket, och Skatteverket i uppdrag av regeringen att främja offentlig förvaltnings förmåga att använda artificiell intelligens (AI) i syfte att stärka Sveriges välfärd och konkurrenskraft.[2] DIGG, är

[1] Europeiska kommissionen, Förslag till Europaparlamentets och rådets förordning om harmoniserade regler för artificiell intelligens (rättsakt om artificiell intelligens) och om ändring av vissa unionslagstiftningsakter, 21.4.2021, COM(2021) 206 final.

[2] Regeringen (21 juni 2021) ”Uppdrag att främja offentlig förvaltnings förmåga att använda artificiell intelligens”. Diarienummer: I2021/01825.

det tänkt, ska samordna myndigheternas arbete, som rapporteras löpande med slutredovisning senast den 20 januari 2023. Uppdragsbeskrivningen hänvisar till den nationella inriktning och den ekonomiska nytta (om 140 miljarder årligen) som DIGG redogjorde för i en rapport från januari 2020 (se vidare avsnitt 2.3). Ett huvudsakligt behov som lyfts i sistnämnda rapport återkommer i uppdragsbeskrivningen för de fyra myndigheterna, nämligen att ta fram en AI-guide för offentlig förvaltning. Guiden är enligt uppdraget tänkt att beskriva de steg en verksamhet behöver ta för att använda sig av AI, inklusive hur AI ska struktureras och tillgängliggöras. Avsikten är att en sådan guide ska anpassas till relevanta *internationella rekommendationer och riktlinjer* för AI-området och utgå från *hållbar AI* i enlighet med den nationella inriktningen för AI. Regeringsuppdraget kan därmed ses bidra i ljuset av EU-kommissionens förslag till rättsakt om AI och en vidare principiell diskurs, ofta formulerad i termer av etisk styrning av AI-tillämpning. Uppdraget sätter därmed fingret på en bredare utveckling vi vill uppmärksamma här: hur AI och databeroende maskininlärning medför metod- och processförändringar för den offentliga förvaltningen. Här väcks därmed en rad styrningsfrågor i den ständigt pågående dialektiken mellan lagstiftningsbehov och tekniska landvinningar.[3]

Samtidigt har sålunda styrningen på AI-området kommit att ta sig uttryck i mjuka former av reglering, genom framtagandet av etiska riktlinjer och värdegrundsbaserade ställningstaganden hos såväl globala företag som internationella sammanslutningar och stater.[4] Hur ser då dessa internationella rekommendationer och riktlinjer ut, vilken typ av medvetenhet och kunskap kring risker med AI bygger dessa på, och hur har myndigheterna hittills influerats av dessa? Vi avser i detta kapitel att teckna de mest centrala idéer om styrningen av AI som framkommit i internationella riktlinjer de senast få åren, med särskilt fokus på EU. Mycket av detta idéinnehåll har också funnit sin väg in i nyssnämnda förslag till en rättsakt om AI, vilket vi därmed också analyserar delar av.

[3] För en diskussion om samspelet mellan teknikutveckling och lagstiftning på AI-området, se Larsson, S. (2021) ”AI in the EU: Ethical Guidelines as a Governance Tool”, i Bakardjieva Engelbrekt, Leijon, Michalski & Oxelheim (red.) *The European Union and the Technology Shift*. Cham, Switzerland: Palgrave Macmillan.

[4] Jfr Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center Research Publication (2020-1).

1.2 Syfte och disposition

Syftet med detta kapitel är att i) ge en kortfattad överblick över hur AI används i svensk offentlig förvaltning, ii) beskriva vilka utmaningar som forskningen pekar på och iii) sammanfatta hur den vidare principiella styrningsdiskussionen ser ut på europeisk nivå.[5] Det betyder att vi analyserar aspekter av EU-kommissionens förslag till rättsakt om AI jämfört med aspekter som lyfts fram längs vägen fram till förslaget – vilka kommit till uttryck i en rad centrala publikationer på AI-området – och ett axplock av svenska myndigheters rapportering av hur AI påverkar offentlig förvaltning. Vi diskuterar här transparens som en av de mest centrala myndighetsfrågorna kopplade till AI och automatiserat beslutsfattande i offentlig förvaltning, inte minst indikerat av kommissionens lagförslag.[6]

Artikelns upplägg är som följer. Först, vi ser ett behov av att börja med en diskussion kring AI-definitionen. Detta eftersom begreppsdefinitionen är helt central för både reglering och myndigheternas teknikutveckling samtidigt som ”AI” är ett begrepp som är påfallande svårt att definiera med exakthet, och dessutom rörligt i sin betydelse över tid.

Därefter, i avsnitt 2, går vi igenom ett axplock av myndighetsrapporter som är mest relevanta för kapitlets syfte. En del av de etiska och rättsliga frågor som pekas ut av både AI-expertgruppen och EU-kommissionen fångas upp av till exempel rapporter från DIGG. En sådan är *Främja den offentliga förvaltningens förmåga att använda AI*, som i början av 2020 bland annat pekade ut behovet av en ändamålsenlig rättsutveckling.[7] En annan sådan är en rapport som tagits fram i samverkan mellan DIGG och Lantmäteriet under 2020 som utforskar hur det offentligas processer och rutiner kan effektiviseras genom automatisering, med bibehållen eller ökad kvalitet, samtidigt som transparensen bakom beslutsfattande bibehålls eller ökar.[8] I rapporten presenteras bland annat ett koncept på

[5] Arbetet med denna artikel har i huvudsak finansierats av Vinnova i projektet Ramverk för Hållbar AI (diarienummer 2019-00848) som har koordinerats av AI Sustainability Center. Vi vill också tacka Jockum Hildén för värdefulla kommentarer på tidigare utkast.

[6] För en inomrättslig analys av hur AI och automatiserat beslutsfattande förhåller sig till framförallt förvaltningsrätt och dataskydd, se Ledendal, J & Larsson, S. (kommande) Artificiell intelligens – rättsliga ramar för automatiserat beslutsfattande i offentlig förvaltning.

[7] Främja den offentliga förvaltningens förmåga att använda AI. Delrapport i regeringsuppdraget I2019/01416/DF || I2019/01020/DF (delvis).

[8] DIGG och Lantmäteriet (2020) *Testa ny teknik för automatisering inom offentlig förvaltning*. I2019/03237/DF.

en förtroendemodell för automatisering i offentlig förvaltning, med fokus på transparens. Detta avseende hur ett system för automatiserade åtgärder är uppbyggt genom tydlighet i form av en deklaration av det automatiserade systemets beståndsdelar.

De etiska och rättsliga frågor som diskuteras på AI-området har ofta sin bas i en forskningsbaserad medvetenhet som vi redovisar i korthet i avsnitt 3, med särskilt fokus på transparensfrågorna. Transparens, tillsammans med frågor om ansvar och rättvisa, är också något som utgör centrala teman för unionens policyutveckling på AI-området, vilket vi redogör för i avsnitt 4. På AI-området pågår som nämnt en betydande utveckling kring styrning (*governance*) av tillämpad AI, i termer av etiska riktlinjer och principiella dokument från både statliga aktörer, civilsamhället, EU-kommissionen och företag.[9] En del av denna utveckling är relevant för de rörelser på området som vi undersöker i denna artikel, och tas därför upp i urval. Mest centralt för kontexten i detta kapitel är EU-kommissionens arbete med AI-frågor så som de tagit sig uttryck genom inrättandet av en AI-expertgrupp (HLEG) som under 2019 och 2020 publicerade åtminstone fyra betydande rapporter, där *Etiska riktlinjer för tillförlitlig AI*[10] haft särskilt märkbar påverkan på nationella strategier i Europa.[11] Dessa publikationer tar i mycket sitt avstamp i kommissionens s.k. AI-strategi från 2018,[12] och har gjort tydliga avtryck i den vitbok som kommissionen publicerade i februari 2020.[13] Vitboken angav indikationer i förhållande till den riskbaserade ansats som återkom i modifierad version i förslaget på rättsakt om AI, vilket vi återkommer till nedan.

I det avslutande avsnitt 5 summerar vi överblicken över AI i offentlig förvaltning i relation till de internationella rekommendationer och riktlinjer vi analyserat, med fokus på de frågor som är mest relevanta för den

[9] Se exv. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399, och Larsson, S. (2020) On the Governance of Artificial Intelligence through Ethics Guidelines, *Asian Journal of Law and Society* 7(3): 437–451.

[10] AI HLEG (2019a) *Etiska riktlinjer för tillförlitlig AI.*

[11] Larsson, S., Ingram Bogusz, C., & Andersson Schwarz, J. (red.) (2020) *Human-Centred AI in the EU. Trustworthiness as a strategic priority in the European Member States*. Brussels: European Liberal Forum.

[12] Europeiska kommissionen, Meddelande från kommissionen till Europaparlamentet, Europeiska rådet, rådet, Ekonomiska och sociala kommittén och Regionkommittén om artificiell intelligens för Europa, 25.4.2018, COM(2018) 237 final.

[13] Europeiska kommissionen, *Vitbok – Om artificiell intelligens – en EU-strategi för spetskompetens och förtroende*, 19.2.2020, COM(2020) 65 final.

nära förestående utvecklingen. DIGG har också nyligen pekat på att det i offentlig förvaltning finns en stor osäkerhet när det gäller digitalisering av ärendeprocesser och automatiserat beslutsfattande. Myndigheten konstaterar att det finns ”olika uppfattningar om de rättsliga förutsättningarna och det saknas tydlighet kring exempelvis frågor gällande offentlighetsprincipen, rättssäkerhet och vad som anses vara god offentlighetsstruktur vid automatiserade förfaranden”.[14] Detta är exempel på utmaningar som kräver mer analys och studier framöver, i takt med att automatiserat beslutsfattande och AI-understödda metoder blir vanligare inom offentlig sektor.[15]

1.3 Kort om AI-definitionens undflyende karaktär

Trots den uppmärksamhet som AI och maskininlärning fått i både media och europeiskt policyarbete råder det inte någon strikt konsensus kring hur AI bäst bör definieras. Definitionen är samtidigt central för de rapporter som försöker mäta om eller hur AI tillämpas eller kalkylera ett monetärt värde på möjlig nytta, exempelvis genom effektivisering av offentlig förvaltning. Hur man definierar avgör helt enkelt vilka värden man summerar, och därmed vilken nytta man ser. Definitionen är vidare också naturligtvis helt central för regleringen av AI. Hur man definierar har direkt bäring på vilka aktiviteter och processer som beläggs med förbud eller kravställningar.

En rad definitioner har lanserats inom såväl forskning som i myndighetsrapporter, och en stor utmaning ligger i att det rör sig om ett dynamiskt och föränderligt fält både gällande vad man konceptuellt betecknar som ”AI”, men också i termer av att teknik- och metodutvecklingen snabbt ökar i kapacitet, kvalitet och precision i sig.[16]

EU-kommissionen angav i ett strategiskt dokument från april 2018 en definition som pekade på hur AI ”avser system som uppvisar intelligent beteende genom att analysera sin miljö och vidta åtgärder – med viss grad

[14] DIGG (2021) Rättsligt stöd till offentlig förvaltning avseende digitalisering Delrapport: Beskrivning av behovet. I2021/00288 Dnr: 2021–164, s. 8.

[15] Jfr Ledendal och Larsson (kommande).

[16] Jfr diskussion i Larsson, S. (2021) ”AI in the EU: Ethical Guidelines as a Governance Tool”, i Bakardjieva Engelbrekt, Leijon, Michalski & Oxelheim (red.) *The European Union and the Technology Shift*. Cham: Palgrave Macmillan.

av självständighet – för att uppnå särskilda mål.”[17] Denna definition förekommer även som utgångspunkt i DIGG:s rapport från januari 2020, och SCB:s överblick från november 2020, som vi återkommer till nedan.

Komplexiteten i begreppsapparaten ledde dock den 52-hövdade AI-expertgrupp som tillsattes av kommissionen 2019 till att föra fram en utvecklad och tämligen mångfacetterad definition i ett specifikt definitionsdokument.[18]

> Artificiella intelligenssystem (AI-system) är programvarusystem (och eventuellt även hårdvarusystem) som har konstruerats av människor och som, när de får ett komplext mål, agerar i den fysiska eller digitala dimensionen genom att uppfatta sin omgivning via datainsamling och att tolka insamlade strukturerade eller ostrukturerade data, resonerar om den kunskap eller behandlar den information som härletts ur denna data och beslutar om den bästa åtgärd eller de bästa åtgärderna som ska vidtas för att uppnå det fastställda målet. AI-system kan använda symboliska regler eller lära sig en numerisk modell. De kan också anpassa sitt beteende genom att analysera hur den omgivande miljön har påverkats av deras föregående åtgärder.

Det finns således olika aspekter att ta fasta på i en definition, där det som pekas ut av AI-expertgruppen är a) AI-*system*, som är b) mänskligt konstruerade, c) målstyrda, d) databeroende, och därmed även anpassningsbara och reaktiva med e) någon grad av agens.

Det kan härmed konstateras att det även finns en särskild utmaning i förflyttningen mellan AI som en kreativ forskningsdisciplin till ett regleringsbegrepp.[19] Det vill säga, att även om en mer rörlig och flytande definition haft sina syften för utvecklingen av metoder som maskininlärning, neurala nätverk och datorseende, så finns det regleringsmässiga utmaningar med ett så pass undflyende begrepp när det används för att styra

[17] Med tillägget ”AI-baserade system kan vara helt programvarubaserade och fungera i den virtuella världen (t.ex. röstassistenter, bildanalysprogram, sökmotorer, tal- och ansiktsigenkänningssystem), eller inbäddas i hårdvaruenheter (t.ex. avancerade robotar, självkörande bilar, drönare eller applikationer för sakernas internet)”. Europeiska kommissionen, 25.4.2018, COM(2018) 237 final.

[18] AI HLEG (2019) En definition av AI: Viktigaste förmågor och vetenskapliga discipliner. https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

[19] Larsson, S. (2021) ”AI in the EU: Ethical Guidelines as a Governance Tool”, i Bakardjieva Engelbrekt, Leijon, Michalski & Oxelheim (red.) *The European Union and the Technology Shift*. Cham, Switzerland: Palgrave Macmillan.

och reglera användningen hos företag och myndigheter. Det påverkar som nämnt också utvärderingar och bedömningar av metodernas potential, vilket vi ska se nedan, för exempelvis användningsområden inom offentlig förvaltning. Det kan därmed också konstateras att AI-expertgruppen och dess verksamhet i huvudsak befinner sig på en generell nivå, medan många av de konkreta utmaningarna behöver kontextualiseras till specifika sektorer, till exempel till svensk offentlig förvaltning, eller undersektorer inom förvaltningen – det är exempelvis stor skillnad på triagering vid akutsjukvård, fördelning av försörjningsstöd i en kommun eller användning av ansiktsigenkänning i en gymnasieskola – för att man ska kunna teckna dess utmaningar tydligare.

Även den föreslagna AI-rättsakten är naturligtvis helt avhängig en reglerad definition, vilket vi återkommer till nedan. Dock, innan vi ser till den internationella utvecklingen kring medvetenhet och styrning av AI, så följer en belysning av AI-användning inom svensk offentlig förvaltning.

2 Hur används artificiell intelligens i offentlig förvaltning?

I detta avsnitt ges en kortfattad överblick över hur AI har hanterats strategiskt i Sverige, och visar på ett urval av rapporter som indikerar på hur AI används i offentlig förvaltning, samt några möjliga utvecklingslinjer.

2.1 En nationell inriktning och en AI-agenda

Den svenska *Nationell inriktning för artificiell intelligens* som publicerades av regeringen 2018 får sägas vara det som mest liknar en officiell strategi, även om det inte är det.[20] Vissa bedömare kallar det en nationell strategi,[21] intressant nog, som därmed ses som tämligen tunn jämfört med övriga nordiska strategier.[22]

[20] Vilket bland annat påpekas i Ek, I. (2021) *AI-politik för konkurrenskraft*. Rapport AU 2022:02:01. Myndigheten för tillväxtpolitiska utvärderingar och analyser (Tillväxtanalys).

[21] Exempelvis EU-kommissionens granskning, se Van Roy, V. (2020) *AI Watch – National strategies on Artificial Intelligence: A European perspective in 2019*, EUR 30102 EN, Publications Office of the European Union, Luxembourg, doi:10.2760/602843, JRC119974.

[22] Robinson, S.C. (2020) ”AI policy in the Nordics. Pledging openness, transparency and trust, while expressing readiness to apply AI in society”, i Larsson, S., Ingram Bogusz,

Ett antal intressenter har dock under ledning av RISE tagit fram ”AI-agendan för Sverige”, med ett 25-punktsprogram, som publicerades den 23 februari 2021 med näringsminister Ibrahim Baylan närvarande. Inte heller detta är dock en officiell strategi. Arbetet har bedrivits i sex olika spår varav offentlig verksamhet och myndigheter är ett. Agendan efterlyser exempelvis AI-rådgivare och förändringsledare för offentlig sektor, på kommunal och regional nivå, samt att AI-satsningar inom offentlig sektor behöver koordineras och drivas mot en vision.

Det finns i AI-agendan en vision om ett ”mer automatiserat och datadrivet samhälle”, ett tydligt fokus på att bryggan och samverkan mellan näringsliv och offentlig sektor behöver utvecklas, stärkas och förenklas. När det gäller forskning och utbildning kan här särskilt lyftas fram ett uttryckt behov av mer satsningar på inter- och multidisciplinär forskning med AI i fokus som kan driva både disciplinär utveckling och AI-forskningen framåt överlag. Inom ramen för detta bör, enligt agendan,[23] bland annat

- universiteten skapa tvärvetenskapliga utbildnings- och forskarskoleprogram;
- forskningsfinansiärer prioritera tvärvetenskapliga utlysningar; samt
- AI definieras som ett eget forskningsområde hos forskningsfinansiärer

En tämligen omfattande del av agendan redogör för rättsliga utmaningar med osäkerhet, exempelvis kring just skydd av personuppgifter. Vidare efterlyser agendan att det juridiska ansvaret i ett alltmer automatiserat beslutsfattande behöver klargöras, vilket inte minst är en viktig fråga för offentlig sektors användning av AI.[24]

2.2 Nordiska ministerrådet om nordiska kommuner

När det Nordiska Ministerrådet under 2019 intervjuade ett urval av nordiska kommuner för att bedöma statusen och vilka frågor som ansågs viktiga, konstaterade man bland annat att användningen mest befann sig på ett teststadium även om det fanns en del tillämpningar i bruk.[25]

C., & Andersson Schwarz, J. (red.) (2020) *Human-Centred AI in the EU. Trustworthiness as a strategic priority in the European Member States*. Brussels: European Liberal Forum.

[23] RISE (2020, s. 4).

[24] RISE (2021, s. 5).

[25] Nordiska Ministerrådet (2019) *Nordiske kommuners arbeid med kunstig intelligens*.

Rapporten visar på två aspekter som var viktiga för den kommunala AI-utvecklingen, där den ena har med relationen mellan offentligt och privat att göra, och den andra med hur kommunens data ofta är organiserat i interna silos. Rapporten pekar särskilt ut betydelsen av tillit i termer av att det finns en risk för att den påverkas på ett negativt sätt om kommunerna inte är tillräckligt transparenta i sin utveckling.[26] Nordiska ministerrådet rekommenderar att kommunerna regelbundet utbyter erfarenheter, genomför forsknings- och utvecklingsprojekt samt utvecklar nordiska etiska riktlinjer. Sistnämnda betonade just vikten av tillit, och poängterade nordiska likheter i värderingar, förvaltningskultur och medborgarnas förtroende för det offentliga.[27]

2.3 DIGG om att främja AI i offentlig förvaltning

DIGG beskriver i en rapport från januari 2020 möjliga värden på upp mot 140 miljarder kronor årligen att omfördela inom offentlig sektor, om AI:s förtjänster kan nyttjas fullt ut. Uppskattningen inkluderar inte kostnader för ett införande – och själva övergången kan förväntas innehålla en rad utmaningar i sig som innebär kostnader – och att göra den här typen av uppskattningar bygger alltid på en rad antaganden som kan vara mer eller mindre robusta, men huvudpoängen kring möjlig nytta är tydlig. En del av svårigheten ligger, återigen, i hur man definierar AI och därmed särskiljer det från exempelvis annan automation i beslutsfattandeprocessen. Definitionen av AI, som vi såg ovan är svår att konkretisera, utgår i DIGG:s rapport från motsvarande i EU-kommissionens Samordnade plan från 2018.[28]

DIGG pekar bland annat på hur AI skulle kunna nyttjas för att möta samhällsutmaningar som:

- en ökad och mera jämlik tillgång till en högkvalitativ hälso- och sjukvård trots vikande ekonomiska förutsättningar och en åldrande befolkning

[26] Nordiska Ministerrådet (2019, s. 36).

[27] Vilket undersöks av Robinson, S. C. (2020). Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society*, *63*, 101421.

[28] Europeiska kommissionen, Meddelande från kommissionen till Europaparlamentet, Europeiska rådet, rådet, Ekonomiska och sociala kommittén och Regionkommittén om samordnad plan om artificiell intelligens, 7.12.2018, COM(2018) 795 final.

- en ökad tillgång till mer jämlik och högkvalitativ utbildning som utgår från den studerandes förutsättningar, trots vikande ekonomiska förutsättningar och lärarbrist
- en offentlig förvaltning som i än större utsträckning än idag försvarar principen om lika behandling och löpande synliggör och åtgärdar upptäckt diskriminering, i syfte att värna tilliten för det offentliga

Rapporten lyfter fram dels rättsliga utmaningar i termer av rättsosäkerhet, personuppgiftsbehandling, tillgång till data från andra, säkerhets- och molnfrågor, samt transparens- och insynsfrågor (bilaga F), dels etiska utmaningar i form av transparens, partiskhet och ansvar (bilaga E). Baserat på sin utredning föreslår DIGG bland annat att ett kompetenscenter med expertis inom AI behöver utvecklas, och att rättsliga förutsättningar skapas för att underlätta försöksverksamhet. Det sistnämnda är en fråga som lyfts av flertalet utredningar,[29] vilket bidragit till att regeringen i januari 2021 uppdrog åt DIGG att tillhandahålla rättsligt stöd till offentlig förvaltning avseende *digitalisering*, vilket ju är en än vidare terminologi. DIGG:s rapport från 2020 föreslår också framtagandet av en ”AI-guide”, vilket också kom att ingå i det förnyade uppdraget som nämns i detta kapitels inledning.

2.4 SCB om AI i Sverige

Statistikmyndigheten SCB fick under 2019 i uppdrag av regeringen[30] att genomföra en kartläggning av användningen av artificiell intelligens inom det svenska näringslivet och offentlig sektor samt inom universitet och högskolor. Man utgick, precis som DIGG ovan från den definition som EU-kommissionen angav i sin strategi från 2018. I den rapport som publicerades i november 2020 konstateras att 10,2 % av verksamheterna inom offentlig sektor uppger att de använt AI i någon form i sin verksamhet under 2019.[31] Utvärderingen konstaterar att anställdas kompetens, utbildning eller erfarenhet utgör det största hindret för användning av AI, och ekar därmed ett av de behov som EU-kommissionens AI-expert-

[29] Jfr Digitaliseringsrättsutredningens slutbetänkande, Juridik som stöd för förvaltningens digitalisering (SOU 2018:25); Kommunutredningens slutbetänkande, Starkare kommuner – med kapacitet att klara välfärdsuppdraget (SOU 2020:8); och Öppna datautredningens delbetänkande Innovation genom information (SOU 2020:55).

[30] Dnr 2019/01964/D.

[31] SCB, Statistiska centralbyrån (2020) *Artificiell intelligens i Sverige*.

grupp propagerar för i termer av ”datafärdigheter” (”data literacy”), eller algoritmkunnighet.

2.5 Riksrevisionen om automatiserat beslutsfattande hos tre myndigheter

När Riksrevisionen under 2020 granskade beslut om föräldrapenning från Försäkringskassan, beslut om årlig inkomstbeskattning av privatpersoner från Skatteverket och beslut om körkortstillstånd från Transportstyrelsen användes förvisso en smalare definition på automation än den AI-definition som DIGG, SCB såväl som EU-kommissionens AI-expertgrupp använder sig av.[32] Riksrevisionen poängterar rentav att inga av de automatiserade besluten för granskning fattas med hjälp av AI.[33] Det konstateras dock att statliga myndigheters *automatiserade* beslutsfattande har lett till ”ökad effektivitet och att grundläggande rättssäkerhetsaspekter har till viss del har förbättrats”.[34] I likhet med några av inspelen ovan konstateras dock problem med att det saknas en tydlig och läsbar dokumentation av den automatiserade handläggningen, det vill säga ett slags transparensproblem. Myndigheterna följer heller inte i tillräckligt hög grad upp huruvida automatiserade beslut blivit korrekta, vilket kan beskrivas som ett behov av bättre spårbarhet, vilket vi återkommer till nedan.

Med denna mer strikta definition av automation konstaterar Riksrevisionen att automatiserat beslutsfattande till viss del har förbättrat grundläggande rättssäkerhetsaspekter, i termer av att även den manuella hanteringen blivit mer enhetlig, vilket bidrar till en ökad likabehandling. Riksrevisionen anser dock att arbetet med automatiserade beslutsprocesser kan struktureras bättre och ge bättre förutsättningar för effektiva, rättssäkra och korrekta automatiserade beslut.

[32] Riksrevisionen (2020:22) *Automatiserat beslutsfattande i statsförvaltningen – effektivt, men kontroll och uppföljning brister*. Stockholm.

[33] Riksrevisionen (2020:22), s. 4.

[34] Ibid.

2.6 DIGG och Lantmäteriet om förtroende och transparens

Lantmäteriet och DIGG utredde i samverkan under 2020 aspekter av transparens och tillit i relation till AI i offentlig sektor. Uppdraget var att utforska hur det offentligas processer och rutiner kan effektiviseras genom automatisering, med bibehållen eller ökad kvalitet, samtidigt som transparensen bakom beslutsfattande bibehålls eller ökar.[35] Utredningen tog fram ett koncept på en s.k. förtroendemodell som samspelar väl med kommissionens och AI-expertgruppens betoning på transparens och spårbarhet för AI-system. Förtroendemodellen syftar till transparens avseende hur ett system för automatiserade åtgärder är uppbyggt genom att sträva efter

- tydlighet i form av en deklaration av det automatiserade systemets beståndsdelar och dess förmåga att utföra uppgifter, eller kompetens, på ett riktigt, rättssäkert och effektivt sätt och
- en strukturerad, säker och öppen logg över systemets och komponenternas identiteter och versioner, innehåll med koppling till specifika ärenden eller åtgärder som utförts

Detta skulle kunna vara en möjlig utvecklingsväg för AI och automatiserat beslutsfattande i offentlig förvaltning där utvecklingsprocess och användningsområden tydliggörs, vilket är en kritisk punkt för den offentliga förvaltningens AI-användning som vi ovan har konstaterat i relation till både smarta städer och andra delar av offentlig förvaltning. Lantmäteriet och DIGG ser att förtroendemodellen för automatiserat beslutsfattande med AI ”kan vara en viktig del för att bibehålla förtroende och tillit till den svenska offentliga förvaltningen men också till den europeiska och globala digitaliserade offentliga förvaltningen”[36] och rekommenderar att det utvecklas och utökas.

Denna förtroendemodell återkommer så också i ovan nämnda regeringsuppdrag som ett flertal myndigheter fick från regeringen i juni 2021.[37] Enligt uppdraget ska den utvecklas ”för att möjliggöra att den

[35] DIGG och Lantmäteriet (2020) *Testa ny teknik för automatisering inom offentlig förvaltning*. I2019/03237/DF.

[36] DIGG och Lantmäteriet (2020), s. 3.

[37] Regeringen (21 juni 2021) ”Uppdrag att främja offentlig förvaltnings förmåga att använda artificiell intelligens”. Diarienummer: I2021/01825.

kan användas som ett frivilligt ramverk för offentlig förvaltning".[38] Rapporten nämner även möjligheten med ett nationellt öppet AI-register, som underlag för hur AI och automatiskt beslutsfattande används inom offentlig förvaltning, men konstaterar att det behöver utredas mer.

3 Forskningsbaserad medvetenhet om AI-baserade risker

Många av de rättsligt relevanta frågorna för styrningen av AI och automatiserat beslutsfattande har vuxit fram i takt med teknologiska möjligheter och forskningsbaserade insikter om både möjligheter och risker. För att förstå vilka vägval som gjorts av till exempel EU-kommissionen och dess AI-expertgrupp behöver man se vilken underliggande medvetenhet och kunskap om relationen mellan AI och samhälle som vuxit fram genom denna typ av forskning. Det är också denna underliggande kunskap som även kan bistå med att vägleda svensk offentlig sektor, vid alla de detaljerade val som behöver göras när AI-metoder implementeras och utvärderas.[39] Det föranleder att en viss överblick ges även i detta avseende.

3.1 Medvetandegörande studier, och principiella ramverk

Det finns en rad exempel från framför allt amerikansk offentlig förvaltning som har problematiserats och studerats på ett sätt som lett till ökad medvetenhet kring risker för diskriminering. Ett sådant är det s.k. COMPAS-systemet, ett system som används i vissa domstolar för bedömning av återfallsrisk hos dömda gärningspersoner. Systemet bedömdes dock i en granskning på felaktiga grunder vara mer benäget att se högre risk hos gärningspersoner av afro-amerikansk bakgrund.[40]

Ansiktsigenkänning är ett annat fält som varit särskilt omdebatterat, både ur integritetsskyddsperspektiv såväl som utifrån diskriminerings- och s.k. bias-perspektiv, vilket inte minst är tydligt gällande kommissionens

[38] Regeringen (21 juni 2021), s. 3.

[39] Jfr DIGG (2020), kap. 8 och bilagor E och F.

[40] ProPublica (2016). "Machine bias", av Angwin, J., Larson, J., Mattu, S., & Kirchner, L.; jfr Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

förslag till rättsakt om AI.[41] Två amerikanska forskare vid MIT analyserade i en studie publicerad 2018 tre kommersiella system för ansiktsigenkänning. De kunde konstatera att samtliga hade sämre precision för kvinnliga ansikten och personer med mörk hy.[42] Något som även påvisats i system som ska upptäcka fotgängare i trafiken.[43] En typ av brist på precision som också tycks ha lett till högre andel felaktiga arresteringar av mörkhyade i USA,[44] och att flera av teknikleverantörerna upphört med att erbjuda den här typen av AI-lösningar för polisiära ändamål. Liknande farhågor har visat sig med diskriminering i system för platsannonser[45] och digital marknadsföring,[46] vilket också kan ställas mot studier som visar hur svårt det är att granska användandet av AI-system och algoritmer för prediktion som utvecklas av privat sektor åt den offentliga förvaltningen.[47]

Denna typ av forskning har lett till insikter om att flervetenskapliga miljöer behövs, och etablerandet av internationella vetenskapliga konferenser, där en av de mer profilerade är FAccT (tidigare FAT).[48] Utvecklingen och medvetenheten kring risker i samspelet mellan AI-teknologier och samhälle syns också i hur traditionella vetenskapliga AI-konferenser

[41] Den europeiska dataskyddsstyrelsen (EDPB) och datatillsynsmannen har till exempel varnat för riskerna med fjärrbiometrisk identifiering av individer i publik miljö, och har i en gemensam åsiktsförklaring anmodan ett allmänt förbud mot det, se EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Se här även Liane Colonnas bidrag i denna volym: The AI Regulation and Higher Education: Preliminary Observations and Critical Perspectives.

[42] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91), PMLR.

[43] Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. *arXiv preprint* arXiv:1902.11097.

[44] The New York Times (24 juni 2020) ”Wrongfully Accused by an Algorithm in what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit”, av Kashmir Hill.

[45] Datta, A., Datta, A., Makagon, J., Mulligan, D. K., & Tschantz, M. C. (2018, January). Discrimination in online advertising: A multidisciplinary inquiry. In *Conference on Fairness, Accountability and Transparency* (pp. 20–34). PMLR.

[46] Latanya Sweeney. Discrimination in online ad delivery. *Commun. ACM*, 56(5):44–54, May 2013.

[47] Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. Yale JL & Tech., 20, 103.

[48] ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT): https://facctconference.org.

tämligen nyligen även börjat adressera etik och samhällsfrågor.[49] För svenskt vidkommande har en inriktning mot humaniora och samhälle i det 10-åriga, AI-inriktade, forskningsprogrammet WASP-HS utmärkt sig.[50]

I linje med detta har även utvecklingen tagit fart kring principiella ställningstaganden, där en rad aktörer från både näringsliv, offentlig sektor och civilsamhället har tagit fram etiska eller rättighetsbaserade riktlinjer för AI-utveckling och -tillämpning.[51] När en forskargrupp vid ett schweiziskt forskningsuniversitet under 2019 analyserade ett stort antal AI-riktlinjer identifierade de inte mindre än 84 dokument som innehöll etiska principer eller riktlinjer för AI.[52] En stor majoritet av dessa, 88 procent, hade publicerats från 2016 till april 2019. De undersökta dokumenten uppvisade konvergens kring de fem etiska principerna för (1) transparens, (2) rättvisa, eller det svåröversatta ”fairness”, (3) icke-skadlighet, (4) ansvar och (5) integritet. De noterade dock också att det verkar finnas stora skillnader i hur dessa principer tolkas, varför de anses viktiga, vilken fråga, domän eller vilka aktörer de relaterar till och hur de ska implementeras.

3.2 Transparensens förtjänster och mångfacetterade betydelse

I principiella ställningstaganden kring att tillämpa AI är behovet av transparens en av de vanligast utpekade principerna. I nämnda kartläggning av riktlinjer pekade 73 av de 84 källorna på just behovet av transparens.[53] En forskningsgenomgång visar dock att begreppet är mångfacetterat, inte minst med tanke på hur olika intressen såsom tillsynsbehov, användarnas informationsfärdigheter (”litteracitet”), företagshemligheter och säkerhetsfrågor behöver balanseras för en teknologi som i sig själv kan ha inne-

[49] Som i att AAAI / ACM utvecklat ett spår om ”Artificial Intelligence, Ethics and Society”.

[50] The Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society: https://wasp-hs.org.

[51] Jfr Fjeld et al. (2020).

[52] Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2.

[53] Jobin, Ienca & Vayena (2019), p. 391.

boende förklarbarhetsproblem.[54] Även EU-kommissionen har i en rapport om säkerhets- och ansvarighetsfrågor, som publicerades samtidigt som vitboken i februari 2020, pekat ut bristen på transparens, och vad de kallar en ”black box effect” som en utmaning för rättslig efterlevnad och ansvarsutkrävande.[55]

Vad det är som behöver vara mer transparent behöver därmed också diskuteras i förhållande till olika aspekter av ett AI-system. Till exempel listar en rapport om ”algoritmisk transparens”, från europaparlamentets forskningsservice, sju åtskilda punkter om vad som behöver åtgärdas. En av dessa riktar sig specifikt till algoritmerna i sig, medan de andra sex behandlar frågor om data, mål, resultat, efterlevnad, inflytande och användning.[56] EU-kommissionens AI-expertgrupp, som vi återkommer till nedan, pekar ut förklarbarhet som en av fyra grundläggande etiska principer att respektera, och transparens som ett av sju krav som AI-system bör uppfylla. Utmaningar med s.k. ”black box”-algoritmer, menar expertgruppen, behöver särskilt uppmärksammas.[57] De lyfter fram behovet av spårbarhet, förklarbarhet och kommunikation. Sistnämnda inkluderar både individer som interagerar med AI-system eller nås av dess utfall, men även de som yrkesmässigt interagerar med AI-system. Detta, menar vi, är särskilt relevant för området för det här kapitlet, det vill säga där handläggare i den offentliga förvaltningen kommer att interagera med exempelvis beslutsstödsystem och samtidigt ansvara för myndighetsbeslut som de fattar under påverkan av dessa stödsystem.

Aspekter av transparens av intresse för offentlig förvaltning är också hur sådan öppenhet ska säkerställas i relation till privata utförare eller utvecklare av system som de senare kan vilja hävda ska vara ”proprietära”, det vill säga deras immateriella egendom eller utgöra företagshemlighe-

[54] Larsson, S. (2019) The Socio-Legal Relevance of Artificial Intelligence, ”Law in an Algorithmic World”, Special Issue of *Droit et Société*. 103(3): 573–593; Larsson, S. & Heintz, F. (2020) Transparency in artificial intelligence, *Internet Policy Review* 9(2): 1–16.

[55] Europeiska kommissionen, Rapport från kommissionen till Europaparlamentet, rådet, Ekonomiska och sociala kommittén om Konsekvenser för säkerhet och ansvar när det gäller artificiell intelligens, sakernas internet och robotteknik, 19.2.2020, COM(2020) 64 final, s. 6–7.

[56] Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). A governance framework for algorithmic accountability and transparency (Study No. PE 624.262) Panel for the Future of Science and Technology, Scientific Foresight Unit (STOA), European Parliamentary Research Service.

[57] AI HLEG (2019a) s. 14.

ter.[58] Larsson och Heintz har analyserat transparensbegreppet i relation till AI utifrån sju intressen som delvis kan vara motstående.[59] Dessa är:

1. Förklarbarhet och utmaningar av s.k. black box-karaktär för AI-system;
2. Rättsliga aspekter av ägande eller företagshemligheter;
3. Behovet av att undvika missbruk eller ”gaming” (som en oönskad konsekvens av öppenhet);[60]
4. Användarnas litteracitet och kunskapsnivå;
5. Språklig eller metaforisk *översättning* av matematiska samband, exv. i användaravtal eller beslut;[61]
6. Marknadskomplexitet, exempelvis i så kallade dataekosystem;
7. Individuellt distribuerat utfall som en tillsynsutmaning.[62]

Avslutningsvis kan man konstatera att AI-transparens både kan avse förståelsen och förklarbarheten hos individuella beslut (utfall), vilket därmed kan ha olika adressater,[63] men också olika aspekter av hur hela

[58] Vilket studeras inom ramarna för s.k. smarta städer av ovan nämnda Brauneis & Goodman (2018), och i förhållande till alltmer automatiserade, datadrivna marknader av den amerikanske rättsvetaren Frank Pasquale (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

[59] Larsson & Heintz (2020), framförallt utvecklat i Larsson (2019).

[60] Vilket ur ett förvaltningsperspektiv i en rapport från Inspektionen för socialförsäkringen (ISF) förs fram av Försäkringskassan i relation till hur transparenta de bör vara med vilka variabler som ingår i hur de genomför riktade kontroller; Inspektionen för socialförsäkringen (2018) *Profilering som urvalsmetod för riktade kontroller. En granskning av träffsäkerheten, effektiviteten och rättssäkerheten i Försäkringskassans modeller för riskbaserade kontroller*. Rapport 2018:5, se exv. s. 86.

[61] Vilket är direkt relevant för automatiserade myndighetsbeslut. Hur de skrivs är rimligen en avgörande faktor för hur den underliggande behandlingen förstås. Vi återkommer till detta i kap. 4 nedan, inte minst i relation till dataskyddets krav på att ge ”meningsfull information om den bakomliggande logiken”, beskrivet i 4.5.

[62] Här avses s.k. personalisering i storskaliga automatiserade system, exv. gällande riktad reklam, och tillsynsmyndigheternas metodutmaning med att kunna granska eventuella otillbörligheter i detta, se exv. Larsson, S. (2018) Algorithmic Governance and the Need for Consumer Empowerment in Data-driven Markets, *Internet Policy Review* 7(2):1–12.

[63] Att AI-systems förklarbarhet behöver relateras till olika adressater (”audiences”) är tydligt uppmärksammat i den s.k. förklarbarhetsforskningen (xAI), jfr Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.

AI-systemet eller beslutsstödet i sig fungerar. Sistnämnda kan i en förvaltningskontext till exempel handla om krav på ”AI-register”, med en typ av öppenhet som finns exempel på i Helsingfors och Amsterdam,[64] eller en typ av ovan nämnda förtroendemodellkort, som Lantmäteriet tagit fram tillsammans med DIGG under 2020.

4 AI i EU: etiska riktlinjer och styrning

I det här avsnittet behandlas hur EU:s policyarbete inom AI-området har utvecklats sedan 2018, genom flera betydelsefulla publikationer, se Figur 4.1 för ett urval: (1) en AI-strategi publicerad i april 2018;[65] (2) en anmodan publicerad i december 2018, riktad till medlemsstaterna att utveckla AI-strategier till mitten av 2019;[66] tillsättande av en expertgrupp (AI HLEG) som bland annat publicerat de (3) Etiska riktlinjer för tillförlitlig AI i april 2019[67] som fått ett relativt stort genomslag i europeiska AI-strategier;[68] efterföljande (4) policy- och investeringsrekommendationer för tillförlitlig artificiell intelligens,[69] samt den (5) vitbok som EU-kommissionen publicerade i februari 2020.[70] Dessa är alla av intresse för att bättre förstå det förslag till rättsakt om AI som kommissionen publicerade i april 2021 och som högst troligt även kommer att ha avsevärd betydelse för regleringen av AI i offentlig sektor. De två avslutande rapporterna från AI-expertgruppen är också intressanta och av olika karaktär: där den ena är en (6) utvecklad självvärderingsmetodik för de som implementerar eller har implementerat AI-system i sin verksamhet,[71]

[64] Analyserat av teknikfilosof Luciano Floridi, som också var medlem i AI HLEG, Floridi, L. (2020). Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki. *Philosophy & Technology*, 33(4), 541–546.

[65] Europeiska kommissionen, 25.4.2018, COM(2018) 237 final.

[66] Europeiska kommissionen, 7.12.2018, COM(2018) 795 final.

[67] Publicerade samtidigt som Europeiska kommissionen, Meddelande från kommissionen till Europaparlamentet, rådet och Ekonomiska och sociala kommittén om att skapa förtroende för människocentrerad artificiell intelligens, 8.4.2019, COM(2019) 168 final.

[68] Larsson, S., Ingram Bogusz, C., & Andersson Schwarz, J. (red.) (2020) *Human-Centred AI in the EU. Trustworthiness as a strategic priority in the European Member States*. Brussels: European Liberal Forum.

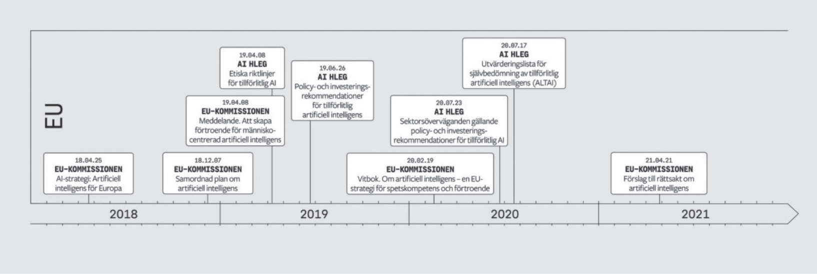
[69] AI HLEG (26 juni 2019) *Policy- och investeringsrekommendationer för tillförlitlig artificiell intelligens*. Bryssel.

[70] Europeiska kommissionen, 19.2.2020, COM(2020) 65 final.

[71] AI HLEG (2020) Utvärderingslista för självvärdering av tillförlitlig artificiell intelligens (ALTAI) [författarnas översättning].

och den andra handlar om (7) överväganden för specifika sektorer, vilket innefattar offentlig sektor.[72] Dessa är baserade på tidigare policy- och investeringsrekommendationer som har bearbetats i workshops under våren 2020 med inbjudna experter och ”stakeholders” (intressenter).

I det följande redogör vi kortfattat för 3–7 i syfte att bättre teckna dels den grund av ”internationella rekommendationer” som DIGG m.fl. myndigheter har att ta hänsyn i inom ovan nämnda uppdrag,[73] dels för att bättre teckna betydelsen av AI-rättsakten för svensk offentlig sektor.


EU
18.04.25
EU-KOMMISSIONEN
AI-strategi: Artificiell intelligens för Europa
18.12.07
EU-KOMMISSIONEN
Samordnad plan om artificiell intelligens
19.04.08
EU-KOMMISSIONEN
Meddelande. Att skapa förtroende för människocentrerad artificiell intelligens
19.04.08
AI HLEG
Etiska riktlinjer för tillförlitlig AI
19.06.26
AI HLEG
Policy- och investeringsrekommendationer för tillförlitlig artificiell intelligens
20.02.19
EU-KOMMISSIONEN
Vitbok. Om artificiell intelligens – en EU-strategi för spetskompetens och förtroende
20.07.23
AI HLEG
Sektorsöverväganden gällande policy- och investeringsrekommendationer för tillförlitlig AI
20.07.17
AI HLEG
Utvärderingslista för självbedömning av tillförlitlig artificiell intelligens (ALTAI)
21.04.21
EU-KOMMISSIONEN
Förslag till rättsakt om artificiell intelligens
2018
2019
2020
2021

Figur 4.1: Policyutveckling mot tillförlitlig AI i EU, i översättning och bearbetning från Larsson et al., 2020.

4.1 Etiska riktlinjer för tillförlitlig AI

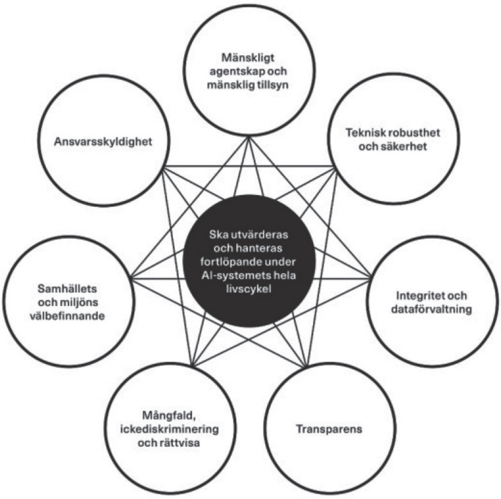
HLEG’s Etiska riktlinjer för tillförlitlig AI publicerades i april 2019 som en följd av det uppdrag som kommissionen givit gruppen. Riktlinjerna består av fyra nivåer:

i. en ram som anger att tillförlitlig AI är laglig, etisk och robust;
ii. etiska grunder för tillförlitlig AI som finns i respekten för mänsklig autonomi, förebyggande av skada, rättvisa och ”förklarbarhet”;
iii. sju krav för att realisera tillförlitlig AI, se Figur 4.2, samt
iv. en utvärderingslista, direkt organiserad enligt dessa sju krav. Denna utvärderingslista testades under andra halvåret 2019 och publicerades i en uppdaterad version i juli 2020 som ”utvärderingslista för självvärdering av tillförlitlig AI (ALTAI)”, avsedd för självutvärderingsändamål.

[72] AI HLEG (2020) *Sektoröverväganden gällande policy- och investeringsrekommendationer för tillförlitlig artificiell intelligens* [författarnas översättning].

[73] Regeringen (21 juni 2021) ”Uppdrag att främja offentlig förvaltnings förmåga att använda artificiell intelligens”. Diarienummer: I2021/01825.

Nivå iii, se Figur 4.2, tycks hittills ha fått mest betydelse för förståelsen av utmaningar i europeiska nationella AI-strategier, vilket är tydligt bland annat i den polska[74] och den norska.[75] Intressant nog är vissa av dessa krav redan tydligt reglerade i rättsordningen, däribland dataskydd och kravet på ickediskriminering. Andra krav, som det om transparens, tycks helt centralt för tillförlitlighetsfrågorna, men är i sin breda betydelse inte lika tydligt reglerat för implementeringen av AI-system i alla sektorer. För den offentliga förvaltningen är dock till exempel den öppenhet som dataskyddsrätten kräver samt rätten till partsinsyn helt relevanta reglerade aspekter av transparens.[76]


Mänskligt agentskap och mänsklig tillsyn
Teknisk robusthet och säkerhet
Integritet och dataförvaltning
Transparens
Mångfald, ickediskriminering och rättvisa
Samhällets och miljöns välbefinnande
Ansvarsskyldighet
Ska utvärderas och hanteras fortlöpande under AI-systemets hela livscykel

Figur 4.2: Sju krav för realiseringen av tillförlitlig AI, från AI HLEG 2019, bearbetad i svensk översättning i Larsson, S. (2020) ”AI i EU: etiska riktlinjer som styrmedel”.

[74] Söderlund, K. (2010) ”AI policy in Poland. Ethical considerations already at the core”, i Larsson, Ingram Bogusz & Andersson Schwarz, red. (2020) *Human-Centred AI in the EU. Trustworthiness as a strategic priority in the European Member States*. Brussels: European Liberal Forum.

[75] af Malmborg, F. (2020) ”AI policy in Norway. Looking to the future and harmonised with the EU”, i Larsson, Ingram Bogusz, & Andersson Schwarz, red. (2020) *Human-Centred AI in the EU. Trustworthiness as a strategic priority in the European Member States*. Brussels: European Liberal Forum.

[76] Se Ledendal & Larsson (kommande).

4.1.1 Policy- och investeringsrekommendationer

Policy- och investeringsrekommendationerna var AI-expertgruppens andra leverans och publicerades den 26 juni 2019. Den omfattar 33 huvudpunkter (med många fler delpunkter) uppdelade i åtta grupper med rekommendationer om

a) mänskligt bemyndigande och skyddsaspekter,
b) transformationen av privat sektor,
c) offentlig sektor som katalysator för hållbar tillväxt,
d) forskningskapacitet,
e) datahantering och infrastrukturella frågor,
f) utbildningsfrågor,
g) styrnings- och regleringsfrågor, och
h) finansieringsfrågor.

Rekommendationen är detaljerad och ger vägledning för många väldigt olika sektorer och ämnen. Av störst relevans här är c och g. Expertgruppen betonar vikten av att använda de etiska riktlinjerna för tillförlitlig AI i offentlig förvaltning (12.1). Den avhandlar även rekommendationer om att individer ska kunna begära att få ”interagera” med en mänsklig handläggare om handläggningen krånglar och medför betydande påverkan på individen (9.2, se även 12.2). Detta ekar dataskyddsrättsliga frågor (se även 27.5 för explicit referens till GDPR). Expertgruppen pekar bland annat på frågor av bredare digitaliseringskaraktär, där data bör finnas digitalt (10.1), och kompetensfrågor i termer av ”datafärdigheter” för myndigheter (10.2).

De pekar också på betydelsen av offentlig upphandling (11.1–3), vilket är ett område vi också ser som särskilt centralt för frågor om AI i offentlig förvaltning, men inte fokuserar i detta kapitel. Intressant nog efterlyser de även granskningsverktyg som kan upptäcka bias och oegentligheter i myndigheternas beslutsfattande (12.3). AI-expertgruppen efterlyser bland annat en översyn av rådande reglering (28.1), utvecklandet av granskningsmekanismer för AI-system (29.4), och en riskbaserad regleringsinriktning (26.1), vilket också fokuseras i kommissionens vitbok.

4.2 Kommissionens vitbok

När vitboken publicerades den 19 februari 2020 åtföljdes den av en rapport om säkerhets- och ansvarseffekterna av AI, IoT och robotik (samt en europeisk datastrategi).[77] Som anges i vitboken är många av de frågor som de etiska riktlinjerna för tillförlitlig AI pekar på redan reglerade, till exempel inom dataskydd och antidiskriminering. Rapporten om säkerhet och ansvar diskuterar konsekvenserna av autonomi och självlärande funktioner hos AI-produkter, särskilt när det gäller riskbedömning. Detta är uppenbarligen av relevans för begreppet *människocentrerad AI*. Dessutom påpekas den brist på transparens och ”black box-effekt” som vissa AI-system kan ha på beslutsprocessen som ett verkställighets- och ansvarsproblem.

Definitionsfrågan, som vi diskuterade ovan, konstateras även här vara central för regleringsfrågorna i termer av att för ett nytt rättsligt instrument måste definitionen vara ”tillräckligt flexibel för att kunna anpassas till den tekniska utvecklingen och samtidigt vara tillräckligt exakt för att ge den rättssäkerhet som krävs”.[78]

Vitboken består av två huvudblock baserade på begreppet ”ekosystem”; ett om excellens och ett om tillit. Det betyder att det finns en dubbelhet i strategin: att undersöka möjligheterna å ena sidan – kopplade till efterfrågan på forskning, samarbete mellan medlemsstaterna, innovation och ökade investeringar – och riskerna eller utmaningarna å andra sidan – för att säkerställa tillförlitlighet, ansvar och säkerhet. Kommissionen konstaterar att lagstiftningsramen kan behöva utvecklas rörande exempelvis:

- effektiv tillämpning och verkställighet av befintlig EU- och nationell lagstiftning. Det vill säga att befintlig lagstiftning i många fall är ändamålsenlig men utmanas i dess implementering. Kommissionen påpekar särskilt bristen på transparens, som gör det svårt att identifiera och bevisa eventuella överträdelser;
- begränsningen av produktsäkerhetslagstiftningens tillämpningsområde som gäller produkter och inte tjänster, och därför i princip inte tjänster baserade på AI-teknik;
- föränderligheten hos AI-system, till exempel för produkter som baseras på maskininlärningsberoende programuppdateringar;
- ansvarsfördelning på olika platser i en försörjningskedja.

[77] Europeiska kommissionen, 19.2.2020, COM(2020) 64 final.

[78] Europeiska kommissionen, 19.2.2020, COM(2020) 65 final, s. 18.

- behov av utveckling av säkerhetsbegreppet, relaterat till exempelvis cybersäkerhet.

Av särskild relevans, och även föremål för debatt, är den föreslagna riskdefinitionen, eftersom den används för att identifiera behovet av framtida reglering. Kommissionen föreslår i vitboken en kumulativ ansats där den dels pekar ut högrisksektorer, till exempel hälso- och sjukvård, transport, energi och delar av den offentliga sektorn, och dels att AI-tillämpningen inom sektorn i fråga används på ett sådant sätt att betydande risker sannolikt uppstår.

Riskdefinitionen har dock fått kritik, eftersom vissa typer av risker inte nödvändigtvis uppstår i högrisksektorer, exempelvis med hänvisning till plattformsmarknader med riktad marknadsföring och sökmotorer, som därmed skulle räknas som AI-system som, trots sina risker, klassificeras som låg risk enligt kommissionens definition.[79] Den tyska regeringen har till exempel krävt att det riskklassificeringssystemet som föreslås i vitboken revideras, vilket i viss mån också kom att göras i det efterföljande förslaget till rättsakt om AI, från en binär ansats till en som har flera nivåer (se nedan).[80]

4.3 Sektoröverväganden och självvärdering

Som nämnt, AI-expertgruppens policy- och investeringsrekommendationer bearbetades under våren 2020 och publicerades med inriktning mot mer konkreta sektorer i slutet av sommaren 2020, varav offentlig förvaltning var en.[81] Expertgruppen uttrycker här behovet av att ”AI-möjliggjorda e-förvaltningstjänster” bör åtföljas av adekvata arrangemang när det gäller ansvarighet och spårbarhet, för att möjliggöra efterhandsverifiering, och hänvisar här även till rätten till god förvaltning, som stadgas i artikel 41 i EU:s stadga om grundläggande rättigheter. Expertgruppen ser även några särskilda behov, exempelvis att vidta åtgärder för att främja data- och algoritmkunnighet (notera att ”datafärdigheter” – ”data

[79] Dignum, V., Muller, C. & Theodorou, A. (2020). First Analysis of the EU Whitepaper on AI. ALLAI; se även Liane Colonnas bidrag i denna volym i relation till högre utbildning och AI.

[80] Die Bundesregierung (2020). Stellungnahme der Bundesregierung der Bundesrepublik Deutschland zum Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen.

[81] AI HLEG (2020) *Sektorsöverväganden*.

literacy" – har breddats) inom den offentliga förvaltningen, samt att utveckla AI- och datastrategier inom alla relevanta grenar av statlig förvaltning. I augusti 2020 publicerades också ALTAI, det vill säga expertgruppens slutgiltiga version av en utvärderingslista för tillförlitlig artificiell intelligens, som pekar på en allt viktigare utvärderingsutveckling som vi dock av utrymmesskäl inte redogör för mer här. Listan följer den sjudelade inriktning för realisering av tillförlitlig AI som först publicerades i de etiska riktlinjerna, här kompletterade med en rad frågor.

4.4 Förslag på ny AI-lagstiftning

Den 21 april 2021 publicerade kommissionen ett förslag till en förordning som syftar till att harmonisera regleringen av artificiell intelligens (rättsakt om AI). Man vill bland annat säkerställa att AI-system som släpps ut och används är säkra och är förenliga med befintlig lagstiftning om de grundläggande rättigheterna och unionens värden. Eftersom den föreslagna regleringen är i form av en förordning blir den till alla delar bindande och direkt tillämplig i varje medlemsstat och ska inte genomföras i svensk rätt, även om vissa kompletteringar kan komma att behöva göras. Den föreslagna rättsakten har naturligtvis inte uppstått i ett politiskt eller kunskapsmässigt vakuum, vilket den ovan beskrivna utvecklingen visar, men relationen till redan befintlig rätt, exempelvis på dataskyddsområdet, väcker en rad frågor. Här finns dock bara utrymme för en kort översikt, med fokus på det som är allra mest relevant för offentlig förvaltning, varför vi enbart redogör för det riskbaserade angreppssättet, hur centralt transparens är och hur man i rättsakten har hanterat AI-definitionen.

4.4.1 Risknivåer som styrande mekanism

Det riskbaserade angreppssättet känns igen från vitboken, men har utökats till att inkludera flera kategorier, från förbud av AI-system med oacceptabel risk, till de med hög risk som kan vara tillåtna under vissa restriktioner och krav, de som innebär låg risk som medför transparensförpliktelser, samt de med minimal risk utan krav. Det ställs krav på både tillhandahållare och användare. I korthet presenteras:

- *Förbud* mot AI-system med oacceptabel risk, som i) använder "subliminala metoder" för att manipulera en persons beteende på ett sätt som innebär skada; ii) utnyttjar sårbarheter hos en specifik (utsatt) grupp för att väsentligt förändra beteendet som innebär eller sannolikt kan innebära (fysisk eller psykisk) skada; iii) profilerar eller klassificerar personers trovärdighet baserat på beteende eller personlighetsdrag om åtgärderna leder till skadlig eller ogynnsam behandling som är obefogad eller oproportionerlig i förhållande till beteendet; iv) biometriskt identifierar personer på distans i realtid på offentliga platser (med vissa möjliga undantag). I den här kategorin finns möjliga applikationer för offentlig sektor framförallt inom det polisiära arbetet, som kommer leda till gränsdragningsutmaningar och debatt.

- AI-system som innebär *hög risk* kan tillåtas om de genomgår en granskningsprocess av en behörig offentlig aktör och beviljas en CE-märkning. Denna ska bland annat innebära en adekvat nivå av transparens kring hur AI-systemet fungerar och dokumentation av systemets funktion för tillsyn och säkerställandet av användning av data som är av hög kvalitet. AI-rättsakten ställer också informations- och transparenskrav för hur AI-system som är avsedda att interagera med fysiska personer ska utformas och utvecklas. Det föreslås också att varje medlemsstat ska utse eller etablera minst en tillsynsmyndighet som ska ansvara för att säkerställa att de nödvändiga förfarandena följs. När förslaget publicerades presenterades också bedömningar av vilken typ av användningsområden som kunde ingå i högriskklassificeringen, och påfallande mycket av offentlig förvaltning fördes fram, varför föreslagna utvärderingar och kravställanden blir relevant för många av de beslutsprocesser som finns inom denna sektor.[82] Noterbart är även att kommissionen föreslås få rätt att genom delegerade akter vid behov lägga till ytterligare AI-system som ska definieras som högrisk, vilket gör att det finns en möjlig rörlighet över tid i vilka tillämpningar som träffas av regleringen.
- AI-system med *begränsad risk* kan ha transparenskrav – till exempel att en mänsklig användare ska veta att det är ett AI-system som interaktionen sker med – och system med minimal risk regleras inte enligt rättsakten.

Rättsakten anger straffavgifter, vars storlek beror på vilket brott mot förordningen som skett. Användningen av förbjudna AI-system ska leda till böter på upp till 30 000 000 euro eller upp till 6 procent av föregående

[82] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

års globala omsättningen om det gäller ett företag vars omsättning är högre än det stadgade bötesbeloppet.

Vi kan konstatera att AI-rättsaktsförslaget är omfattande, varför också bedömningen av konsekvenserna alla dess komponenter i det här stadiet är svårt. Infrastrukturdepartementet skickade förslaget på remiss den 21 maj 2021, och bad om svar senast den 24 juni 2021 och även om man kan misstänka att många av remissinstanserna har haft svårt att bilda sig en initierad överblick på så kort tid så finns det några anmärkningar att särskilt notera.[83] Remissinstanser lyfter bland annat fram utmaningar med regelkonflikter eller otydlighet i relation till befintlig lagstiftning (DIGG, IMY). Vissa betonar övergångsproblematiken för redan etablerade AI-system (Arbetsförmedlingen), eller att det inte är tydligt hur skiljelinjen går mellan forskningens behov och användningen på den inre marknaden (bland annat Linköpings universitet och Formas).

4.4.2 Definitionen i den föreslagna AI-rättsakten

Flera av de svenska remissinstanser anmärker på den vida definitionen av artificiell intelligens i kommissionens förslag på AI-rättsakt (bland annat DIGG och Lunds universitet), vilket föranleder att ett särskilt fokus läggs vid den här. I både kommissionens första definition såväl som i AI-expertgruppens omformulerade version, som vi redogjort för ovan, finns ett mått av självständighet eller autonomi med som inte återfinns i förslaget på rättsakt. I huvudsak regleras AI-definitionen i rättsaktens Art. 3(1) som anger att AI-system är:

> programvara som utvecklats med en eller flera av de tekniker och metoder som förtecknas i bilaga I och som, för en viss uppsättning människodefinierade mål, kan generera utdata såsom innehåll, förutsägelser, rekommendationer eller beslut som påverkar de miljöer som de samverkar med.

Denna definition ska läsas tillsammans med bilaga 1, som inkluderar flera begrepp av vid karaktär, exempelvis ”statistiska metoder” och ”optimeringsmetoder”:

a) Metoder för maskininlärning, inbegripet övervakad, oövervakad och förstärkt inlärning, med hjälp av en mängd olika tillvägagångssätt, inklusive djupinlärning.

[83] Regeringskansliet ”Remiss av Europeiska kommissionens förslag till förordning om harmoniserade regler för artificiell intelligens”.

b) Logik- och kunskapsbaserade metoder, inklusive kunskapsrepresentation, induktiv (logisk) programmering, kunskapsbaser, inferens- och deduktionsmotorer, (symboliska) resonemang och expertsystem.
c) Statistiska metoder, bayesisk beräkning, sök- och optimeringsmetoder.

Utmaningen med en alltför vid definition påpekas bland annat av DIGG:s remissvar i termer av de kostnader som medföljer utvärderings- och uppföljningskraven. Vi har i kapitlets inledning konstaterat att ”AI” är ett mångbottnat och föränderligt koncept, vilket föranleder tydliga utmaningar när begreppsapparaten ska styra en reglering. Det finns en risk för att en alltför vid definition leder till att AI-rättsakten får karaktären av en mer allmän teknikreglering. Vi antar därför att definitionen kommer att bli mer strikt i lagstiftningens slutversion.

5 Övergripande slutsatser: Från etik till juridik

Som nämnt i inledningen fick DIGG, Arbetsförmedlingen, Bolagsverket, och Skatteverket i juni 2021 i uppdrag av regeringen att främja offentlig förvaltnings förmåga att använda artificiell intelligens (AI), i syfte att stärka Sveriges välfärd och konkurrenskraft.[84] Uppdraget hänvisade till behovet av en AI-guide, att anpassa efter internationella rekommendationer och riktlinjer, vilket vi kortfattat har redovisat ovan. Vi har även visat på hur vissa myndigheter redan har tagit sig an uppgiften och influerats av de idéer om behov av styrning som många av dessa riktlinjer uttrycker, inte minst gällande transparens, ansvarsfördelning och rättvisefrågor, ofta uttryckta i det något svåröversatta fairnessbegreppet.[85]

Genom att i det här kapitlet ge en överblick över hur AI används i svensk offentlig förvaltning och ställa den utvecklingen i kontrast till både de idéer och utmaningar som forskningen pekar på i den vidare principiella styrningsdiskussionen på europeisk nivå kan vi konstatera att:

- AI-regleringsidéerna har utvecklats mycket över relativt få år i en dialektik med insikter inom en flervetenskaplig forskning, och har lett till en flora av ”mjuk” reglering i form av rekommendationer och etiska

[84] Regeringen (21 juni 2021) ”Uppdrag att främja offentlig förvaltnings förmåga att använda artificiell intelligens”. Diarienummer: I2021/01825.

[85] Jämför exempelvis Larsson, S. (2018) ”Sjyst AI och normativ design”, i Akenine, D. & Stier, J. (red.) *Människor och AI*. Stockholm: AddAI.

riktlinjer. Förslaget på en europeisk AI-rättsakt pekar dock på att en formalisering i termer av en juridifiering av delar av dessa insikter är på gång, med betydande konsekvenser för både AI-utveckling och tillämpning av AI i offentlig förvaltning.

- Definitionen av AI är påfallande svårfångad, konceptuellt, och rörlig, forskningsmässigt, vilket leder till en specifik styrningsutmaning och heterogenitet i internationella rekommendationer såväl som i kommissionens förslag på AI-rättsakt. Detta är också en fråga för bedömningen av betydelsen av AI inom offentlig förvaltning – som finns i spännvidden mellan varianter av regelbaserad automation, som funnits länge, och mer autonomt lärande och prediktiva rekommendationssystem baserade på maskininlärning, som är ett nyare fenomen.
- Flertalet av de svenska myndighetsrapporter som behandlar AI-relaterade frågor, och inte minst kommissionens förslag på AI-rättsakt, lägger tonvikt vid behovet av transparens i tillämpningen av AI-system. Begreppet är dock mångbottnat, och här betonas både behovet av att kunna granska och idka tillsyn över tillämpningar genom exempel på AI-register och förtroendemodellkort, såväl som både tjänstepersonens relation till AI-system – som ju kan ansvara för myndighetsbeslut där AI-system spelat en avgörande roll i en underliggande bedömning – och slutanvändarens förståelse för vem eller vad hen interagerar med.
- I litteraturen pekas bland annat på behovet av att bättre hantera utmaningar med transparens och ansvar i relationen mellan offentlig förvaltning och privat utförare. Här finns intressemotsättningar mellan det allmännas behov av insyn och företagens ägarintresse och att exempelvis hemlighålla konkurrensfördelar. Här kan registerföring säkerställa nödvändig granskningsbarhet. Även upphandlingsområdet ser till exempel ut att spela en särskilt central roll för AI-utvecklingen inom offentlig förvaltning, både gällande granskningsbarhet såväl som för hur nyttan av offentligt insamlade data, som används för att träna upp kommersiella system, också säkerställs komma till det allmännas godo.
- Utvecklingen mot en rättslig formalisering kan vara gynnsam för utvecklingen av AI i offentlig sektor, eftersom bristen på klargörande rättsregler kan vara ett hinder för myndigheter. Vi ser dock här ett behov av fler studier och klargöranden kring hur den statliga värde-

grunden och principer för god förvaltning kan säkerställas även när AI-system används som beslutsstöd eller rentav beslutsfattare i offentlig förvaltning. Ytterst, för att fortsatt säkerställa ett upprätthållande av förvaltningens nödvändiga tillförlitlighet.

Johan Eddebo & Anna-Sara Lind

Artificial Intelligence and Imperceptible Governance via Opinion Formation: Reflections on Power and Transparency from a Cross-Disciplinary Encounter[1]

1 Introduction

In a society characterized by a bourgeoning and increasingly ubiquitous digital infrastructure, hardly any field is left untouched by the increasing reliance on artificial intelligence (AI). This is also true for governance, that is, how governments and organizations steer and control behavior. One aspect of the impact of AI on governance that has received much attention (see for example in this volume the contribution by Markku Suksi) is the incipient use of automated decision-making (ADM) in the public sector. Caution is increasingly recommended due to the risks and issues relating to the introduction AI within present frameworks of governance and decision-making.[2] Examples of risks associated with the use of ADM in the public sector are lack of transparency and effective accountability, which can be partly due to technological and organizational issues, and partly to a lack of clear regulatory provisions on the legislative side.

[1] The authors are grateful for valuable inputs and comments by Ass. Professor Sandra Friberg, PhD Oliver Li and Assistant Prof. Katja de Vries.

[2] See *Dagens Nyheter*. "Myndighetsbeslut måste alltid vara rättssäkra – oavsett om de fattas av människor eller maskiner". Online: https://www.dn.se/ledare/myndighetsbeslut-maste-alltid-vara-rattssakra-oavsett-om-de-fattas-av-manniskor-eller-maskiner/.

All the same, it must be stated that algorithmic systems already exert a significant influence over decision-making processes, both directly and indirectly. Notwithstanding their formal integration in explicit governance, AI is already a substantial factor in terms of the formation of public opinion and political consent,[3] particularly in the conduct of communications and flows of information in digital contexts.[4] Such influence is arguably a factor in all flows of information in digital contexts where algorithms influence the perceptions of the recipient in any way. As a general remark, we distinguish between three levels of AI-systems where mainly two of these are central to the discussion. First, we have the simple algorithms used in such circumstances as data-filtering processes, such as a bit of code that picks out posts with a certain frequency of listed keywords. Then we have the more advanced systems based in evolving or self-learning algorithms basically capable of processing large amounts of data, and then "extrapolating" factors relevant for future decisions. These systems can be characterized by an inherent unpredictability even from the programmer's perspective. Thirdly, there is the hypothetical category of strong AI exhibiting behaviour indistinguishable from that of human agents, in principle able to perform any type of decision-making task. This category will not be addressed in this chapter, however.

This chapter will focus on the implicit governance exercised through AI which arguably already is in place and expanding, and give special attention to a form of complex or layered opacity peculiar to the phenomenon. That is, indirect algorithmic governance effected by e.g. private tech-platforms firstly employ AI systems which are proprietary and inaccessible to external review. In our contribution, we assume a general definition of governance as processes of policy creation involving different actors and networks, which impact upon social formation and the reproduction or establishment of institutions. Secondly, it is very difficult to get a sense of the actual effects of this automated discourse manage-

[3] See SOU 2014:75 pp. 31–32; Young Mie Kim, "Algorithmic Opportunity: Digital Advertising and Inequality in Political Involvement", *The Forum*, 14 (4), 2016.

[4] Cf. Samuel C. Woolley, Philip N. Howard (eds.), *Computational Propaganda*, New York: OUP 2019; Ujué Agudo, Helena Matute, "The influence of algorithms on political and dating decisions". *PLoS ONE* 16(4), 2021. Online: https://doi.org/10.1371/journal.pone.0249454; Riksrevisionen, Automatiserat beslutsfattande i statsförvaltningen – effektivt, men kontroll och uppföljning brister (RiR 2020:22). The shaping of consent via mediatic processes is a contentious topic all by itself, beset by significant issues relating to the principles of rational, unguided and uncompelled democratic deliberation.

ment due to the unpredictability of the underlying algorithms and the lack of access to the platforms' private data on traffic and information flow. We will argue for the importance of mitigation strategies, present a set of viable options, and discuss legislative possibilities. The chapter's outline is as follows. First, in section 2, a brief description of the current situation will be presented. Here we aim to sketch the frames of the issues that we would like to study closer and we explain the arguments as to how AI is in effect exerting de facto governing functions. Thereafter, in section 3, our attention turns to the concept of *double intransparency* and how it is manifested in decision-making processes. In section 4, possible mitigation strategies, such as for example the Artificial Intelligence Act, are discussed but also other legislative options are presented in light of our results (section 5). We then conclude our contribution (section 6) with observations relating to our discussions and findings.

Furthermore, it should be addressed that this chapter has been written jointly by a philosopher of religion and a legal scholar.[5] The dialogue between the authors has been accordingly done across disciplinary borders and is worth reflecting upon. As the attentive reader will see, the disciplinary backgrounds of these two authors do have an impact on how our questions are asked and on how the discussion is framed. In the chapter's concluding part, we will accordingly come back to this and reflect upon our encounter.

Methodologically speaking, philosophy's general approach is to scrutinize the meaning of abstractions, in terms of everything from pure concepts to established social institutions. When the discussion regards a complex social situation like the present one, a useful way to proceed is to then explore the possible theoretical and structural implications of the relevant abstractions. Here one attempts to discern which consequences, applications or developments are likely or even inevitable in principle (or vice versa) when these abstractions are taken to regulate or guide social processes. This discernment is also preferably anchored in supporting empirical data or material expressing the intentions of the institutions and agents involved.

This type of general theoretical overview may come off as naïve from the point of view of jurisprudence or the social sciences, which are more

[5] The authors are part of the national research program WASP-HS for more information see wasp-hs.org) and collaborates in the project Artificial intelligence, democracy and human dignity.

familiar with the details of the complex limitations and possibilities of the social structures involved. The advantage is that this perspective may also afford novel solutions difficult to discern from within these inevitably entrenched specialized disciplines.

In working with this chapter, we have attempted to proceed by first establishing philosophy's more unfiltered speculative suggestions, and then relating them to actual legislation and juridical practice. All through the working process, we have had a continuous dialogue and several meetings exchanging views and learning from each other's field. This is also shown in the text as it to a large extent mirrors the ongoing dialogue between the authors.[6] Our experience is that this multidisciplinary dialogue can add new insights to the respective disciplines involved but also contribute to developing new questions.

2 The current situation – points of departure

In February 2021, Facebook announced that 97% of all "hate speech" was pre-emptively detected and removed by their automated systems before any human had flagged it.[7] Their proactive removal was said to rely upon a complex contextual analysis of language and the communicative setting. The interrelation of text, comments and images was ostensibly taken into account, which was said to enable a high accuracy of the automated decisions. Hate speech is an increasingly prominent concept in the contemporary political discourse, a conceptual construct characterized by a certain ambiguity, which in the case of targeted suppression or censorship efforts adds another level of transparency issues.[8] Correspondingly, Facebook defines hate speech as any type of communication which attacks people in relation to their "protected characteristics", while adding that there is no consensus in terms of exactly what amounts to a transgression in this sense.[9]

[6] Compare with Lind, A-S, *Den offentliga rätten i mångvetenskaplig forskning*, pp. 207 ff.

[7] Schropfer, Mike, "Update on Our Progress on AI and Hate Speech Detection", *Facebook* 2021. Online: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/.

[8] Cf. Brudholm, Thomas, Johansen, Schepelern Brigitte (eds.) *Hate Politics Law*, New York: OUP 2018, p. 5–11.

[9] Richard Allan, "Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?", *Facebook* 2017. Online: https://about.fb.com/news/2017/06/hard-questions-hate-speech/.

These kinds of assurances of successful and effective detection are quite optimistic, particularly in relation to the complexity added by the formal ambiguity of hate speech as a concept, and one ought to be sceptical of the validity of such purported automatic assessments. This level of complex estimation is namely difficult even for human persons. It arguably necessitates a thorough familiarity with the cultures, languages and social settings involved, which renders doubtful the accuracy of such automated flagging which at best can mark patterns of symbolic association from predetermined directives such as lists. Whereas a human can rationally recognize a certain kind of contextually relative speech act and even empathize with the intention of its originator so as to actually understand it, the AI will at best only flag possible and probable associations. Moreover, the human need not rigidly adhere to a fixed set of symbols or list of connections, but can make a reasoned assessment of social interactions and possible breaches of trust and etiquette in a fluid environment.

All this is to say that the AI is likely to err in ways human persons would not, while sometimes also being prone to reinforce human error, such as the aggravation of biases.[10] To this we must add that the proliferation of automated AI review and censorship will literally impact trillions of interactions. We are here in a sense dealing with something akin to the butterfly effect, where a small change in the initial conditions of the algorithm, especially if self-learning, will likely produce immense and unforeseeable effects upon interaction patterns and the flow of information.[11] The potential ramifications are varied and far-reaching, and makes the admittedly massive influence of traditional mass media seem rather primitive and superficial in comparison.

That interference at this scale is likely to be characterized by structurally proliferated errors in judgment is severely problematic. Of greater importance, however, are the potential political effects of this type of automated interventions. Remaining with the example of hate speech suppression, one or the first remarks often made by scholars is that the very definition of the concept is rife with ambiguity and contradiction.[12] The accounts

[10] Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, "Machine Bias", *Propublica* 2016. Online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[11] Robert M. Entman, Nikki Usher, "Framing in a Fractured Democracy: Impacts of Digital Technology on Ideology, Power and Cascading Network Activation", *Communication Theory* vol. 68, no. 2 2018.

[12] Alexander Brown, *Hate Speech Law*, New York: Routledge 2015, p. 4–5.

of hate speech outside the field of law are quite varied, in accordance with the diversity of the fields researching the concept, such as sociology, psychology, linguistics and political science, while already jurisprudence as such exhibits a broad range of competing descriptions.[13] Such ambiguity coupled with large-scale suppressive interventions by private operators of the digital infrastructure renders the politicization of hate speech countermeasures a distinct possibility. As a kind of speech act which must be defined contextually, in relation to its purpose and effects, hate speech as such can often not be unambiguously specified,[14] and the actual stipulations thereof arguably give a great leeway for politicized interpretations.[15]

A related example is found in Facebook's recent push towards stifling that which is designated "political extremism". In practice, algorithms filter out posts and comments according to undisclosed criteria that the user cannot access, and issue warnings to the poster's contacts as well as recipients of the information. The poster is on the other hand not notified of the warnings, which either amount to implying that an acquaintance of the recipient might be "becoming an extremist", or informing the recipient that he or she "may have been exposed to harmful extremist content recently".[16] The obvious consequence that friends and acquaintances of the originator of content designated as politically problematic will be targeted, ostensibly with the purpose of creating an indirect social pressure to counteract such perspectives deemed as politically problematic.

"Extremism" as a concept is characterized by ambiguities far more significant than those regarding hate speech. As an unqualified noun, its function is entirely relative to some implicit norm, and can in practice refer to any kind of political position whatsoever. The extremism decried by Justin Trudeau will very likely differ from the one opposed by Viktor Orbán. However, Facebook reassuringly lets us know that it cooperates with NGOs and academic experts in developing these algorithmic countermeasures against politically undesirable content, so we should be confident that the system will not be abused.[17]

[13] Cf. e.g. Glasser 1994; Vasquez and de las Fuentes 2000; Fraleigh and Tuman 2011.

[14] Stavros Assimakoupoulos, Fabienne H. Baider, Sharon Millar, *Online Hate Speech in the European Union*, Cham: Springer 2017, p. 3–4.

[15] Cf. Nadine Strossen, *Hate*, New York: OUP 2018, Chs. 4 & 7.

[16] BBC 2021.

[17] Nawab Osman, Adam Burke, "New Resources to counter Hate and Extremism Online". *Facebook* 2021. Online: https://about.fb.com/news/2021/06/new-resources-to-counter-hate-and-extremism-online/.

All the same, the formal and structural problems remain in place, while private authorities independently and without much oversight providing the key definitions for a far-reaching automated discourse formation. This kind of interference, and at the massive scale the interrelated social media corporations operate, arguably amounts to a form of de facto governance in practice, and holds an obvious potential to exert political influence in myriad ways.

The establishment of an AI-based "counter-disinformation" framework by institutions such as Facebook is one particularly clear example.[18] The approach here is to simply suppress politically undesirable content by either removing the posts as such, or demoting them so that they are unlikely to appear in various feeds and will spread little when shared. These processes will in turn affect the visibility of the primary media outlets which to some extent depend upon the tech platforms as infrastructure, and strongly encourage an adaptation of their content in line with that which is selected as acceptable by the custodians of the platform.

This is inevitably going to shape discourses and opinion formation in the public sphere, with potential ramifications for any kind of decision-making that can be influenced by prominent media narratives. Interference of this kind must therefore be considered a type of governance even in accordance with more stringent definitions of the concept, since it actively delineates acceptable public policy and reproduces public consent for preferred positions.[19]

Another example which touches upon the broad range of further possibilities in terms of governance is the tech platforms' proactive "nudging" of users who have simply interacted with politically objectionable content. The concept, coined by Nobel Laureate Richard Thaler, entails the preemptive shaping of an informational environment so as to control the outcome of subsequent decisions.[20] In practice, certain users who have "liked" or otherwise interacted with undesirable content are then targeted with information promoting divergent perspectives or simply negating the content in question. The actual efficacy of these interven-

[18] Linda Slapakova, "Towards an AI-based Counter-Disinformation Framework", *The Hague Diplomacy Blog* 2021; Techcrunch 2021 https://techcrunch.com/2020/05/12/facebook-upgrades-its-ai-to-better-tackle-covid-19-misinformation-and-hate-speech/.

[19] Michael Barnett, Raymond Duvall, *Power in Global Governance*, New York: Cambridge University Press 2005, p. 15.

[20] Richard Thaler, *Nudge: Improving Decisions About Health, Wealth and Happiness*, New Haven: Yale University Press 2008.

tions are supported by research indicating that, at least in certain situations, these types of interference in the dissemination of marginalized narratives successfully discredit the targeted stories about half the time. In other words, after being confronted with a seemingly authoritative correction, half of the users rescinded belief in the addressed narratives.[21] Even minimal nudging has been shown to change the consensus perspective towards a preferred position, and merely simple filtering algorithms rather than complex self-learning algorithmic systems seem sufficient for an effective regulation of opinion formation.[22]

3 The double intransparency of the imperceptible influence of AI over decision-making processes and opinion formation

Transparency is an important aspect of the exercise of power in open societies for several reasons. Most obviously, it is an important part of the rule of law. It safeguards accountability for errors and abuses, which depends on a visible chain of decisions and a clearly established causation. Transparency is naturally also key with regard to a functional democratic influence over the decision-making processes in society, without which there can be no proper open deliberation over governance and the distribution of power. In Swedish constitutional law, transparency is expressed in several ways. It is embraced and promoted in the Freedom of the Press Act where the right to access to official documents is stated.[23] Transparency is also achieved through the constitutional demand that the work of courts and legislative bodies should be done publicly.[24] In the Administrative Procedure Act documentation and motivation of decisions are expressed

[21] Avaaz, "White Paper: Correcting the Record", *Avaaz.com*, 2021. Online: https://secure.avaaz.org/campaign/en/correct_the_record_study/.

[22] Nicola Perra, Luis E. C. Rocha, "Modelling opinion dynamics in the age of algorithmic personalization", *Scientific Reports 9*, 7261, 2019. Online: https://doi.org/10.1038/s41598-019-43830-2.

[23] Chapter 2 Section 1 Freedom of the Press Act. See also the contributions made by Johan Hirschfeldt and Anna-Sara Lind respectively in the anthology Transparency in the future – Swedish Openness 250 years (Lind, Reichel and Österdahl, Eds.), 2017. See also Axberger, pp. 43–44 and Lind (2015).

[24] See for example Chapter 2 Section 11 para. 2 Instrument of Government and Chapter 5 Section 42 the Local Government Act (2017:725).

as a rule and a general demand for all public bodies and those who have been given the right to take administrative decisions.[25]

Indeed, even in an authoritarian society, a minimal level of transparency with regard to the exercise of power is arguably necessary to enable regulatory oversight and to avoid abuse and the entrenchment of corruption, if for no other reason than to maintain basic safety, the efficiency of production, and social cohesion and stability.[26]

And given what is mentioned in the previous section pertaining to the current situation, there's an obvious argument that the influence exercised through algorithms within the framework of digital communication platforms and contemporary media technologies exhibits a complex form of intransparency. This is more specifically what we referred to as "double" intransparency in the introduction, since it pertains to both our inability to access the algorithms as such, as well as the difficulties of reproducing and examining their actual effects.

To begin with, there is no clear and reliable way for the end-user in a social media framework to ascertain whether he or she is being subtly "nudged" by having his information feed or search results altered in relation to an algorithmic assessment of what information is deemed appropriate, or even if the data is being tailored in accordance with the user's traced prior activity. The information being presented to us may have been algorithmically prioritized to the detriment of our access to other information, or with the purpose of discrediting certain narratives and perspectives, and there is no obvious way to determine whether or not this is the case. Neither is it in most instances possible for the user to find out if any of the information he or she has passed on is in any way suppressed or diverted, as in the case of "shadow banning" (although irregular patterns in others' interactions can possibly be a sign). Shadow banning is the practice whereby a user's communication is suppressed, e.g. by downranking his or her content to that it is less visible or almost invisible in others' news feeds. This is naturally also much less conspicuous than the outright banning of a user. Marked and sudden reductions in interactions is the only obvious indication that something like this may have taken place.

[25] Sections 27, 28, 31 and 32 Administrative Procedure Act.

[26] In *Transparency and Authoritarian Rule in Southeast Asia* (London: Routledge 2004), Garry Rodan for instance argues that transparency measures beneficial from an efficiency perspective have been implemented in Southeast Asia while indirectly supporting rather than challenging authoritarian rule.

If we then add to the equation the self-learning aspect of advanced AI, we also have an inherent unpredictability in the situation which essentially precludes full transparency, even from the perspective of those providing the initial programming. In other words, a self-learning decision-making algorithm may effect changes in relation to a fluid information environment which are in principle unpredictable at the outset.

When AI systems influence opinion formation at a very large scale (Google for instance serves almost 4 billion search queries per day),[27] we are faced with a situation of an almost imperceptible influence over popular opinion, over media narratives, and indirectly parliamentary processes. We have no real access to many of the filtering algorithms since they are proprietary (or at least trade secrets) and difficult to reverse engineer, and any user data from which researchers could empirically infer interaction patterns and the presence of bias or active influence is likewise privately held.

Moreover, the influential interactions between the user and something like politically modified search results are transient and short-lived. They generally cannot be recreated after the fact due to the changing information environment, nor registered in any straightforward way, which renders them nearly impossible to audit.

When we then establish the active influencing of popular opinion and political processes using these kinds of tools, a qualitatively new, and quite undetectable form of governance, has essentially been set up. To be sure, an entity like Facebook has only utilized these technologies of control in relation to contentious issues such as hate speech or purportedly false information, yet they also refuse to explicitly repudiate a broader usage:

> Facebook declined to answer a question from Recode about whether it will apply its warnings to other types of misinformation in the future.
>
> For companies like Facebook, it's a lot easier to draw a line in the sand on misinformation about coronavirus topics than around more politically contentious ones, like gun rights, abortion, immigration, or even the 2020 US elections.[28]

[27] Internet Live Stats 2021. Online: https://www.internetlivestats.com/google-search-statistics/.

[28] Shin Ghaffary, "Facebook will start nudging users who have 'liked' coronavirus hoaxes", *Recode* 2020. Online: https://www.vox.com/recode/2020/4/16/21223972/facebook-coronavirus-hoaxes-warning-misinformation-avaaz.

As for Google, the active engineering of search results for political ends has been a debated issue for several years.[29] But notwithstanding the actual scope or character of these kinds of influence, their "benevolent" implementation will normalize structures of intervention and institutionalize these new forms of governance. Interventions targeting fake news and perspectives associated with universally derided groups may well be especially prone to catalysing such a development due to the assent it can plausibly engender. For instance, popular opinion might be quite accepting of the suppression of political views associated with movements such as the radical right, whether or not their connection is accidental.

4 Reflections relating to possible solutions and mitigation strategies

So far, we can conclude that the presence and influence of AI also in private settings is getting more complex. This complexity should however not be understood as impossible to handle. Possible solutions need however to take into account that when it comes to the transnational nature of AI, several jurisdictions interplay and are applicable at the same time. This also means that different traditions, administrative settings, legal cultures and a variety of bodies need to interact with the legislative measures and forms chosen.

One possible approach towards addressing issues of influence and opacity, which at the same time circumvents some of this institutional complexity, could be that we embark upon a major quest to surveil the algorithmic impact of all these private digital media and platforms.[30] This would necessitate impartial surveillance bodies, using methods and tests for gathering information (anonymous of course) in order to achieve a transparent comparison of how these private systems generate different

[29] Cf. John D. McKinnon, Douglas MacMillan, "Google Workers Discussed Tweaking Search Function to Counter Travel Ban", *Wall Street Journal* 2018. Online: https://www.wsj.com/articles/google-workers-discussed-tweaking-search-function-to-counter-travel-ban-1537488472; Kirsten Grind, Sam Schechner et al., "How Google Interferes With Its Search Algorithms and Changes Your Results", *Wall Street Journal* 2019. Online: https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753.

[30] Cf. the Digital Services Act as an example. See also Bernard Rieder: https://policyreview.info/articles/analysis/towards-platform-observability.

political nudges depending upon the users' profiles and online behaviour. This suggestion is also in principle possible to realise using existing legislation, and does not necessitate any newer intrusive or controversial legal measures as long as it links to and respect the realisation of fundamental rights, such as the rights to privacy and/or data protection rules.[31] Its effectiveness, however, seems to be contingent upon the quality of a complex network of supervision, as well as of the effective broadcasting and political reception of this network's reviews and critique.

Another theoretically possible option could be to actually expropriate and make public all data the tech platforms gather and use in order to realise the different search results. In this way, the data resources could eventually become a type of universal big data in the form of a "public commons" accessible to all.[32] Accordingly, the monopolistic situation of a few multinational companies would be undermined and other private entities as well as states would have new opportunities to e.g. create their own search engines open to comparisons focusing on unwanted bias patterns. In this situation, open research could additionally in principle render transparent such "informational influence" that today is exerted with the use of proprietary data banks and digital platforms. A related option would be to create a parallel system of a big data commons that would not expropriate current private entities per se, but that over time would be able to mirror their information resources and thereby both challenge and scrutinize their political and economic influence.

Especially the second option is however not easily realised due to the pluralistic legal landscape of contemporary society. The two models need to be adapted to national and European law (both the European Union as well as the Council of Europe) and respect fundamental and human rights relating to amongst other things, the right to privacy and the right to property. This becomes more difficult as our discussion involves private

[31] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), the European Charter in Fundamental Rights, the Treaty on the European Union, the Treaty on the Functioning of the European Union as well as the relevant legal sources from the Council of Europe.

[32] See for example the proposed Data Governance Act: https://www.consilium.europa.eu/en/press/press-releases/2021/10/01/eu-looks-to-make-data-sharing-easier-council-agrees-position-on-data-governance-act/. See also http://infolegproject.net/call-for-papers-for-workshop-data-and-the-common/.

multinational companies, having rights in their own capacity according to both private and public law. Traces of this can also be seen in the work done by the European Commission. Moreover, the political influence of large multinational corporations would likely hamper any initiatives towards actual expropriation of what in many cases is their key resource and source of income. This could in turn trigger further consolidation from the corporate side, possibly exacerbating present issues of regulatory capture and corporate political influence, prompting these already very prominent private entities to strengthen their ties to legislative processes.

5 Legislative options – current initiatives

There are currently several approaches being discussed pertaining to handling and steering the development of AI both for private and public use. The legal initiatives taken come from several actors of different sorts. Private companies, often international and dominant, have tried to create their own communication strategy for legitimizing the use of algorithms, as we have seen above with the example of Facebook. There are also international organisations drafting soft law documents elaborated by states in dialogue with different expert groups. Online public consultation on these matters attires great attention.[33]

The latest initiative, however, is taken by the European Commission in its Proposal for an Artificial Intelligence Act (AI Act).[34] This proposal followed after intense discussions since the General Data Protection Regulation was enacted in 2016. In the EU, a high-level expert group on AI (HLEG), comprised of 52 experts was established and so was an AI Alliance with 4000 stakeholders, holding a yearly AI Assembly. The European Parliament and the Council explicitly requested a process of regulating AI in 2017 and in the political guidelines 2019–2024 entitled "A Union that strives for more", this was underlined.

[33] See for example the work with the AI Act within the European Union, Proposal for a Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM(2021) 206. See also how the process of enacting the Data Protection Regulation followed a similar procedure before the EU Commission started its work as stated in the EU Treaties. Reichel, Jane & Lind, Anna-Sara, Regulating Data Protection in the EU, I: Perspectives on Privacy, Dörr, Dieter & Weaver, Russell L. (ed.), de Gruyters publisher, 2014, pp. 22–45.

[34] Proposal for a Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM(2021) 206.

The two options suggested in the previous section would not have great success within the framework of the AI Act. In the Proposal, focus is on risk management and a risk-based assessment of the different AI systems is key. The stronger the risk, the more safeguards are demanded. Transparency becomes a core value, according to the Proposal, when we deal with a high-risk AI system. The AI Act is construed in such a way that it imposes legal rules for high risk AI systems while all those delivering AI that is not of high risk need to voluntarily obey to codes of conduct. Further clarification is accordingly needed as to how this should be handled. High risk is decided upon the level of negative effects a certain system has. If it is unacceptably high, the system is forbidden. Some high risk systems might be allowed, however, if there are public security aims legitimising the use of the system. In those circumstances, certain conditions must be met. For the high risk systems, documentation, data governance, transparency and information to users are core conditions that need to be fulfilled. It is also expressly stated that the system needs to have human oversight (Article 14 AI Act).

The AI Act explicitly states that it will not contravene the mechanisms put in place through the GDPR, nor consumer protection, non-discrimination and gender equality expressed in different legislative documents. The risk of overlapping legislative acts is clear. In the AI Act the European Commission tries to handle this through references of giving priority to the GDPR as soon as privacy issues relating to individuals are at stake.[35] In the new administrative structure suggested by the Commission, the European AI Board needs to consult and respect the opinion of the European Data Protection Board.

The complex issue of transparency, legal certainty and surveillance can be illustrated with the following. In the AI Act, market surveillance mechanisms are suggested. As the Act mostly has as its legal basis the internal market rules in the Treaty on the Functioning of the European Union (art. 114), it mainly focuses, as we have seen, on the realisation of free market rules and rights. The market surveillance authorities are suggested to be public bodies. The idea is to have a system of notifications in place so that users inform providers if risks have appeared or if some-

[35] See for example para. 23 in the preamble.

thing is not working properly.[36] The providers must inform the market surveillance authorities if their post-marketing monitoring identifies risks or cases of non-compliance. This mechanism is only designed for the AI systems showing to be of higher risk. Unfortunately, the individual is not given the opportunity to sue a provider or user for not respecting the AI Act. The right to lodge complaints for the individual has not been included in the AI Act and it does not contain any such right for groups either. These basic dimensions would have been a simple and obvious way to strengthen transparency and foster good development in future realisation of AI.

6 Concluding observations

Current and proposed legislative and regulatory approaches request a stringent assessment of the potential harm of AI systems, precautionary measures ensuring the transparency of "high risk systems", as well as effective human oversight. In light of the arguments and analyses presented in this chapter, two main questions follow in relation to these approaches.

1. How can a reasonable distinction of high vs. low-risk systems be made in relation the AI systems in question?

It seems difficult to designate any type of AI system as low-risk when even the most basic types of AI, such as simple filtering algorithms, can have complex and opaque effects in the sense that they have a certain unpredictability in terms of their actual consequences – especially when they are integrated into a fluid informational environment where large numbers of institutions, groups and individuals interact. Moreover, any self-learning algorithm whose operations may impact human individuals in a chaotic environment will by definition have uncertain and possibly far-reaching consequences and would seem to preclude anything akin to a low-risk assessment. To actually gauge the risk of such systems in any meaningful way seems to require an evaluation of their actual operations

[36] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011.

and interaction with their intended context, i.e. much cannot really be said until after the fact of their implementation.

2. What types of precautionary measures can in practice ensure transparency within the current legal frameworks and tools of enforcement?

Ensuring transparency in the face of the difficulties we have discussed is a daunting task. Fines or similar sanctions are rarely effective measures against very large corporate entities, especially if they possess some level of influence over legislative bodies. Relying on voluntary obedience in relation to codes of conduct is also hardly a perdurable solution due to the nature of the modern corporation and the competitive environment in which it operates.

Our suggestions, which to some extent are amenable to the current legislative framework, promise to address these issues by building or promoting new structures for review and surveillance, where also in principle any potentially influential AI system can be placed under scrutiny, enabling a more comprehensive risk assessment before more intrusive measures are taken. Importantly, they at least open the door to meaningfully ask whether these incipient forms of technology and certain applications of them are really desirable from a cost-benefit perspective at an early stage before their full entrenchment in society. In relation to question 1 we would like to state that the AI Act does something resembling an contextual assessment by differentiating between low-risk and high-risk uses. That is a good thing, but it is a problematic contextual assessment in the sense that high-risk and low-risk is a problematic distinction. It is difficult, even impossible, to state that AI systems or application contexts are by definition low-risk. This trend to create "free-zones", where you only have voluntary regulation, might therefore not be desirable. At the same time, this is done in the context of the General Data Protection and in a broader sense one could see the GDPR as a sort of a regulatory safety net also for AI related legal issues. This is also underlined in the AI Act and in the Explanations presented by the commission, for example through the supremacy of the European Data Protection Board.

The AI Act also promotes new precautionary structures for review and surveillance of AI systems which leads us to our question 2. This is done as reliance on voluntary codes of conduct does not perdurable solution due to the nature of the modern corporation and the competitive environment in which it operates. Once again, this means that the regulatory free-zone for low-risk AI systems is not very desirable.

A final reflection in this chapter is a methodological one. As mentioned above in section 1, this text is a methodological encounter between two authors who have different disciplinary backgrounds, philosophy of religion and law respectively. What have been the gains? We believe that the different sources used in this text widen our perspectives and deepen our knowledge. It has also been rewarding to seek fruitful interactions between our different methodological approaches, which although related bring out quite different aspects of the problems under scrutiny. It should however be stressed that the authors do know each other professionally, as they are part of the same research group and have spent time talking to each other cross-disciplines before, which likely has expedited the process of interdisciplinary collaboration as there already is an established heuristic framework in place.

And what can we, respectively, take with us in the process of continued work on AI? Law as it is communicated and construed in different jurisdictions, and at different levels, has become more fragmented and pluralistic than was the case before. This has led to a stronger presence of the constitutional dimensions of law, that in turn embrace the dimensions of power and legitimacy relevant at all levels and a common denominator of all jurisdictions. This is true also for European law and the interplay between European, international and national law. To understand and explain these mechanisms and phenomena in relation to AI is however not the sole task of the legal domain. Law needs to interact and communicate with society in many ways and is mirrored in the changes that occur in society. To explain, visualise and test future challenges is however an endeavour that in itself must by definition be multidisciplinary. This deliberation must invite all parts of civil society as far as possible, in the interest of consolidating the foundations of liberal democracy in this transformative period characterised by complex and diverse challenges thereof.

Bibliography

Preparatory works, Sweden

SOU 2014:75 Automatiserade beslut – färre regler ger tydligare reglering

Riksrevisionen, Automatiserat beslutsfattande i statsförvaltningen – effektivt, men kontroll och uppföljning brister (RiR 2020:22)

European Union

Proposal for a Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM(2021) 206

Literature

Axberger, Hans-Gunnar, Constitutional Responsibility for the Free Flow of Information and Ideas in the Internet Age, in: Lind, Reichel & Österdahl (Eds.), Information and Law in Transition – Freedom of Speech, the Internet, Privacy and Democracy in the 21st Century, Martinus Nijhoff Publicher and Liber, 2015, pp. 43–54

Barnett, Michael, Duvall, Raymond, *Power in Global Governance*, New York: Cambridge University Press 2005.

Brown, Alexander, *Hate Speech Law: A Philosophical Investigation*, New York: Routledge 2015.

Brudholm, Thomas, Johansen, Schepelern Brigitte (eds.) *Hate Politics Law*, New York: OUP 2018.

Fraleigh, D.M. and Tuman, J.S., *Freedom of Expression in the Marketplace of Ideas*, Thousand Oaks, CA: Sage 2011.

Glasser, I. "Introduction", in H.L. Gates Jr. et al. (eds.) *Speaking of Race, Speaking of Sex: Hate Speech, Civil Rights, and Civil Liberties*, New York: New York University Press 1994

Hirschfeldt, Johan, "Free access to public documents – a heritage from 1766", in Lind, Reichel and Österdahl, (eds.) *Transparency in the future – Swedish Openness 250 years*, Ragulka Press, 2017.

Lind, Anna-Sara, Sweden: "Free press as a first fundamental right", in Suksi, M, et al. (Eds.) *First fundamental rights documents in Europe*, Intersentia, 2015

Lind, Anna-Sara, "Freedom of the Press Act – from then to now" in Lind, Reichel and Österdahl, (eds.) *Transparency in the future – Swedish Openness 250 years*, Ragulka Press, 2017.

Lind, Anna-Sara, "Den offentliga rätten i mångvetenskaplig forskning", i Arvidsson, et al. (eds.), *Festskrift till Wiweka Warnling Conradsson*, Jure förlag 2019.

Rodan, Garry, *Transparency and Authoritarian Rule in Southeast Asia*, London: Routledge 2004.

Thaler, Richard, *Nudge: Improving Decisions About Health, Wealth and Happiness*, New Haven: Yale University Press 2008.

Vasquez, M. and de las Fuentes, C., "Hate Speech or Freedom of Expression? Balancing Autonomy and Feminist Ethics in a Pluralistic Society", in M. Brabeck (ed.) *Practicing Feminist Ethics in Psychology*. Washington: American Psychological Association 2000.

Woolley, Samuel C., Howard, Philip N. (eds.), *Computational Propaganda*, New York: OUP 2019.

Other sources

Angwin, Julia, Larson, Jeff, Mattu, Surya & Kirchner, Lauren, "Machine Bias", *Propublica* 2016. Online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Agudo, Ujué, Matute, Helena, "The influence of algorithms on political and dating decisions". *PLoS ONE* 16(4), 2021. Online: https://doi.org/10.1371/journal.pone.0249454

Assimakoupoulos, Stavros, Baider, Fabienne H., Millar, Sharon, *Online Hate Speech in the European Union*, Cham: Springer 2017.

Avaaz, "White Paper: Correcting the Record", *Avaaz.com*, 2021. Online: https://secure.avaaz.org/campaign/en/correct_the_record_study/

BBC, "Facebook tests extremist content warning messages", *BBC News* 2021. Online: https://www.bbc.com/news/technology-57697779

Dagens Nyheter, "Myndighetsbeslut måste alltid vara rättsäkra". Online: https://www.dn.se/ledare/myndighetsbeslut-maste-alltid-vara-ratts-sakra-oavsett-om-de-fattas-av-manniskor-eller-maskiner/, *Dagens Nyheter* 2021.

Entman, Robert M., Usher, Nikki, "Framing in a Fractured Democracy: Impacts of Digital Technology on Ideology, Power and Cascading Network Activation", *Communication Theory* vol. 68, no. 2 2018.

Ghafary, Shin, "Facebook will start nudging users who have 'liked' coronavirus hoaxes", *Recode* 2020. Online: https://www.vox.com/recode/2020/4/16/21223972/facebook-coronavirus-hoaxes-warning-misinformation-avaaz

Grind, Kirsten, Schechner, Sam et al., "How Google Interferes With Its Search Algorithms and Changes Your Results", *Wall Street Journal* 2019. Online: https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753

Internet Live Stats 2021. Online: https://www.internetlivestats.com/google-search-statistics/

Kim, Young Mie, "Algorithmic Opportunity: Digital Advertising and Inequality in Political Involvement", *The Forum*, 14 (4), 2016.

MacMillan, Douglas, McKinnon, John D., "Google Workers Discussed Tweaking Search Function to Counter Travel Ban", *Wall Street Journal* 2018. Online: https://www.wsj.com/articles/google-workers-discussed-tweaking-search-function-to-counter-travel-ban-1537488472

Osman, Nawab Burke, Adam, "New Resources to counter Hate and Extremism Online". *Facebook* 2021. Online: https://about.fb.com/news/2021/06/new-resources-to-counter-hate-and-extremism-online/

Perra, Nicola, Rocha, Luis E. C., "Modelling opinion dynamics in the age of algorithmic personalization", *Scientific Reports* 9, 7261, 2019. Online: https://doi.org/10.1038/s41598-019-43830-2

Schropfer, Mike, "Update on Our Progress on AI and Hate Speech Detection", *Facebook* 2021. Online: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/

Slapakova, Linda, "Towards an AI-based Counter-Disinformation Framework", *The Hague Diplomacy Blog* 2021. Online: https://www.universiteitleiden.nl/hjd/news/2021/blog-post---towards-an-ai-based-counter-disinformation-framework

Malou Larsson Klevhill, Annina H. Persson
& Magnus Strand

Ansvarsfrågor vid algoritmisk handel med finansiella instrument[1]

1 Inledning

Algoritmisk handel med finansiella instrument (en del av fenomenet ”FinTech”) har pågått sedan 1990-talet, med början på de nordamerikanska marknadsplatserna. Den algoritmiska handelns effekter på finansmarknaden, och i synnerhet på dess stabilitet, har diskuterats sedan dess.[2] Händelser såsom den så kallade ”flash crash”[3] som drabbade USA den 6 maj 2010 (där avvikande handelsmönster orsakade av algoritmer medförde en tillfällig nedgång motsvarande ett samlat värde av 1 000 000 000 000 USD[4]) har förstärkt oron och skapat en medvetenhet

[1] *Föreliggande arbete har har erhållit stöd från Marianne och Marcus Wallenbergs stiftelse inom ramen för forskningsprogrammet the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS).* Författarna vill varmt tacka jur. dr Johanna Chamberlain, professor Mattias Dahlberg, jur. dr Andreas Kotsios och jur. dr Katja de Vries för värdefulla synpunkter på tidigare utkast till texten. Som vanligt är alla misstag författarnas egna.

[2] För en översikt se Chaboud, A P, Chiquoine, B, Hjalmarsson, E, och Vega, C, Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market, The Journal of Finance 69(5) (2014) s. 2045. Hur stor del av den sammantagna handeln med finansiella instrument som nu sker genom högfrekvenshandel är svårt att uppskatta eftersom frågan faller sönder i flera olika frågor. Den intresserade rekommenderas att läsa Aldridge, I, och Krawciw, S, Real-Time Risk: What Investors Should Know about FinTech, High-Frequency Trading, and Flash Crashes, Wiley 2017.

[3] Med en ”flash crash” avses plötsligt fallande kurs/kurser på marknadsplatsen.

[4] En kort beskrivning av händelsen finns t.ex. i Bostrom, N, Superintelligens, Fri Tanke 2020, s. 40 f. En utförlig redogörelse lämnades till det amerikanska representanthusets Subcommittee on Capital Markets, Insurance and Government Sponsored Enterprises av

om att det behövs ett system för att hantera den ökade volatilitet som orsakas av högfrekvenshandel. Som vi kommer att illustrera ytterligare nedan behövs ett tydligt system för allokering av rättsligt ansvar när värden förloras genom brister i algoritmiska system för handel med finansiella instrument.

En stor del av denna algoritmiska handel sker med övermänskligt hög frekvens, så kallad algoritmisk högfrekvenshandel eller HFT (high-frequency trading). HFT innebär att användarna kapitaliserar på placeringar av ett stort antal ordrar och snabbt rör sig över flera marknader. Den stora mängden snabba transaktioner är avsedd att genom överläggsen timing i förhållande till pris, kvantitet och så vidare generera värdeökningar genom handel i en hastighet och med en frekvens som inte är möjlig för en mänsklig trader. Bland den algoritmiska handelns andra positiva effekter brukar nämnas att likviditet tillförs marknaden, liksom att handeln blir mer systematisk när den effekt som mänskliga känslor kan ha på handeln uteblir. Nervositet och flockbeteenden som påverkar handel negativt kan alltså undvikas, i vart fall under förutsättning att algoritmen är programmerad att bortse från vissa börsrörelser som orsakas av den mänskliga faktorn. Andra exempel på fördelar som brukar framhållas är låga transaktionskostnader vid handel, samtidig kontroll över flera marknadsplatser, minskad risk för manuella felgrepp när ordrar placeras, och möjligheter att spåra historiska data såväl som data i realtid för att finna en framgångsrik handelsstrategi.

Det finns dock en hel del risker med algoritmisk handel som kan överskugga fördelarna. Som exempel kan nämnas operationella risker, det vill säga problem som skapas av bristfällig teknisk infrastruktur eller av externa händelser såsom ett strömavbrott. Även bristfälligt programmerade eller tränade algoritmer är sårbara för sådana risker. Andra typer av systemrisker kan uppkomma exempelvis av att många likartade algoritmer samtidigt fattar samma beslut. Detta kan ge upphov till dramatiska kurssvängningar på grund av samtidig försäljning och ge upphov till så kallade ”flash crashes”. En annan typ av risk har att göra med förtroendet för det finansiella systemet: den som har tillgång till algoritmisk handel får en konkurrensfördel, medan den enskilda pensionsspararen knappast kan följa med i handeln på jämbördiga villkor. Det kan också uppstå oro

Mary L. Shapiro, ordförande för U.S. Securities and Exchange Commission, och finns tillgänglig på https://www.sec.gov/news/testimony/2010/ts051110mls.htm.

för prismanipulationer såsom ”icebergs”[5] eller att algoritmerna agerar på insiderinformation. Förtroendet för börsernas och myndigheternas förmåga att hantera de potentiella problem som algoritmer kan ge upphov till kan också skadas om man inom finansbranschen och bland allmänheten upplever att tillsynen brister eller att ansvarsstrukturerna kring algoritmisk handel är otillräckliga.

I den här artikeln kartlägger vi därför och diskuterar de ansvarsstrukturer som finns på plats under det gällande regelverket. Vi identifierar fyra olika aspekter som vi uppfattar som problematiska:

1. Regelsystemets komplexitet
2. De tillämpliga definitionernas bristande tydlighet
3. Oklarheter kring marknadsoperatörernas (börsernas) ansvar
4. Det oklara civilrättsliga ansvaret

Vi kommer att behandla dessa nedan. För att ge läsaren en mer fullständig bild kommer vi också att redovisa aspekter som i en svensk kontext är relativt tydliga, som exempelvis regelverket om ansvarsutkrävande genom administrativa sanktioner.[6]

För tydlighets skull kan algoritmisk handel med finansiella instrument ske såväl med hjälp av algoritmer som kan kvalificeras som artificiell intelligens (AI) som av algoritmer som faller utanför AI-begreppet. Som berörs av flera bidrag i den här boken är AI mycket svårt att definiera.[7] För våra syften kan AI förstås som ett automatiserat system vars programmering inte enbart består av ett på förhand givet flödesschema (”decision tree”) utan som självständigt identifierar relevanta beslutsstyrande faktorer, bedömer hur de ska värderas, och drar egna slutsatser. En avancerad AI kan också själv utveckla sin egen programmering utifrån givna mål. I AI-baserad handel med finansiella instrument blir det fråga om att AI:n

[5] Med ”iceberg” avses en situation där en algoritm beordrats att på ett fördolt sätt sälja ett större innehav, genom t.ex. uppdelning i många mindre poster.

[6] I internationell kontext kan det dock vara problematiskt att tillsynsmyndigheter i olika länder har olika inriktning i sin tillsyn och gör olika val av åtgärder. Om detta se Larsson Klevhill, M, och Persson, A H, Supervisory Arbitrage: The Case of Sweden. I: Legal Accountability in EU Markets for Financial Instruments: The Dual Role of Investment Firms, Oxford University Press 2021, s. 184–247.

[7] Jfr den synnerligen tungfotade definition som tagits in i Kommissionen, Förslag till Europaparlamentets och Rådets förordning om harmoniserade regler för artificiell intelligens (Rättsakt om artificiell intelligens), COM(2021) 206 final, föreslagen art. 3(1) med hänvisning till annex 1.

självständigt analyserar de finansiella instrumentens värdeutveckling, fattar beslut om transaktioner, och genomför transaktioner. Samtidigt ska det alltså noteras att algoritmisk handel också kan ske med hjälp av förprogrammerade indikatorer som avgör under vilka förutsättningar algoritmen ska genomföra transaktioner. Någon specialreglering av just AI-baserad algoritmisk handel finns inte, utan båda dessa varianter behandlas här, liksom i regelverken, gemensamt under rubriken algoritmisk handel. Vi vill dock redan här markera att många av de problem vi uppfattat, och som redovisas nedan, kan accentueras när den underliggande tekniken bygger på mer avancerad AI-teknik.

2 Regelsystemet och dess komplexitet

2.1 Översikt: EU-rättslig och svensk normgivning

Kommissionens förslag till förordning om harmoniserade regler för artificiell intelligens innehåller inte heller några skrivningar om algoritmisk handel. Sedan tidigare regleras algoritmisk handel istället genom EU:s gemensamma regelverk om marknader för finansiella instrument. De centrala lagstiftningsakterna på området är de så kallade MiFID II[8] och MiFIR.[9] De relevanta materiella reglerna om algoritmisk handel finns i direktivet MiFID II.[10] Medan MiFIR är en EU-förordning, och således bindande och direkt tillämplig i Sverige utan ytterligare nationella åtgärder,[11] är MiFID II ett direktiv vars regler om algoritmisk handel införlivats i svensk rätt genom ändringar i lagen (2007:528) om värdepappersmarknaden (8 kap. 23–27 §§),[12] nedan kallad VpmL. De materiella

[8] Europaparlamentets och Rådets direktiv 2014/65/EU av den 15 maj 2015 om marknader för finansiella instrument [2014] OT L 173/349.

[9] Europaparlamentets och Rådets förordning (EU) nr 600/2014 av den 15 maj 2015 om marknader för finansiella instrument [2014] OT L 173/84.

[10] Det ska för fullständighetens skull noteras att det i art 26 MiFIR finns en skyldighet för värdepappersföretag att i rapporter till myndighet (i Sverige Finansinspektionen) om genomförda värdepapperstransaktioner också identifiera den algoritm som i förekommande fall beslutat och genomfört en transaktion.

[11] Art 288 FEUF. Medlemsstaterna är enligt EU-domstolens rättspraxis förhindrade att överföra reglerna i en EU-förordning till nationellt antagna rättsregler, eftersom ett sådant förfarande kan skapa ”oklarhet, både vad gäller de tillämpliga bestämmelsernas rättsliga karaktär och tidpunkten för deras ikraftträdande”, mål 39/72 *Kommissionen mot Italien* EU:C:1973:13, p. 17.

[12] Lag (2017:679) om ändring i lagen (2007:528) om värdepappersmarknaden.

lagreglerna kompletteras av ytterligare EU-instrument, som med avseende på algoritmisk handel antagits i form av en delegerad EU-förordning.[13] Den parallella normgivningen på EU-nivå och nationell nivå, med delvis överlappande terminologi, kan vara förvirrande och kräver i vart fall en kort förklaring.

Inom de gränser och på de sätt som följer av EU-fördragen har medlemsstaterna gett EU (eller rättare sagt EU:s institutioner) viss normgivningsmakt. EU utövar huvudsakligen sin normgivningsmakt genom att anta rättsakter. De olika typerna av rättsakter finns definierade i artikel 288 FEUF. Rättsakterna kan dock antas på olika sätt och får därigenom olika placering i den EU-rättsliga normhierarkin. Rättsakter som antas i enlighet med det ordinarie lagstiftningsförfarandet (eller i vissa fall ett särskilt lagstiftningsförfarande) kallas enligt artikel 289 FEUF ”lagstiftningsakter”. Det finns emellertid också möjlighet att inom ramen för en lagstiftningsakt delegera viss normgivningsmakt till Kommissionen (så kallade ”delegerade akter”, artikel 290 FEUF), liksom en möjlighet för Kommissionen (eller i vissa fall Rådet) att, när så krävs, anta rättsakter angående genomförande av lagstiftningsakter (så kallade ”genomförandeakter”, artikel 291 FEUF).[14] Detta kan jämföras med att riksdagen i lag delegerar viss normgivningsmakt till regeringen, respektive en svensk myndighets meddelande av verkställighetsföreskrifter till en överordnad författning. Såväl delegerade akter som genomförandeakter antas vanligen i form av förordningar vilka, som ovan nämnts, är bindande och direkt tillämpliga i samtliga medlemsstater. När det är fråga om EU-normgivning kan alltså EU, såsom i det här fallet, ha antagit en lagstiftningsakt som är ett direktiv, vars regler emellertid fylls ut och kompletteras av regler i till exempel en delegerad akt som är en förordning. Den svenska lagstiftning (VpmL) som införlivar direktivets (MiFID II) regler i svensk rätt kommer alltså att kompletteras av den EU-förordning (MiFIR) som

[13] Kommissionens delegerade förordning (EU) 2017/565 av den 25 april 2016 om komplettering av Europaparlamentets och rådets direktiv 2014/65/EU vad gäller organisatoriska krav och villkor för verksamheten i värdepappersföretag, och definitioner för tillämpning av det direktivet [2017] OT L 87/1. Regler angående dokumentation och rapportering finns dessutom i Kommissionens delegerade förordning (EU) 2017/589 av den 19 juli 2016 om komplettering av Europaparlamentets och Rådets direktiv 2014/65/EU med avseende på tekniska tillsynsstandarder som fastställer organisatoriska krav på värdepappersföretag som bedriver algoritmisk handel [2017] OT L 87/417.

[14] Mer om detta kan läsas i valfri EU-rättslig lärobok, t.ex. Bergström, C F, och Hettne, J, Introduktion till EU-rätten, Studentlitteratur 2014, s. 41 ff.

Kommissionen antagit. Just MiFID II och MiFIR kompletteras av ett femtiotal delegerade akter och genomförandeakter, varav nästan alla är förordningar.[15] En av dessa avser alltså just algoritmisk handel och kompletterar därmed de regler från MiFID II som för svenskt vidkommande återfinns i 8 kap. VpmL.[16]

2.2 Materiellt regelinnehåll

I ingressen till MiFID II uttrycker EU-lagstiftaren oro till exempel över ”att algoritmiska handelssystem ska överreagera på andra marknadshändelser, vilket kan förvärra volatiliteten om det redan finns ett problem på marknaden”, eller i övrigt ”brista på ett sätt som kan skapa en oordnad marknad”.[17] Reglerna om algoritmisk handel är därför utformade för att säkra det finansiella systemets stabilitet genom att se till att de värdepappersinstitut som bedriver sådan handel har en tillräckligt robust organisation och teknisk infrastruktur för sådan handel. 8 kap. 23 § 1 st. VpmL är illustrativ:

> Ett värdepappersinstitut som bedriver algoritmisk handel ska ha effektiva system och riskkontroller som är anpassade för den verksamheten. Systemen och kontrollerna ska säkerställa att institutets handelssystem är motståndskraftiga och har tillräcklig kapacitet, att de omfattas av lämpliga handelströsklar och handelslimiter och att de förhindrar att felaktiga order skickas eller att systemet på annat sätt fungerar så att det kan skapa eller bidra till en oordnad marknad.

Det gäller bland annat också att säkerställa sin driftsäkerhet, så att avbrott inte förekommer (3 st.). Värdepappersinstitut som ägnar sig åt algoritmisk handel ska också anmäla detta till Finansinspektionen och eventu-

[15] En översikt och mera utförlig kommentar finns i Strand, M, The choice of instrument for EU legislation: mapping the system of governance under MiFID II and MiFIR. I: Governing Finance in Europe: A Centralisation of Rulemaking?, Edward Elgar Publishing 2020. Detta är dock ett rörligt mål, så siffrorna i artikeln får i någon mån betraktas som dagsnoteringar.

[16] Det kan tilläggas att den behöriga Europeiska myndigheten, European Securities and Markets Authority (ESMA), också utfärdar vissa vägledande instrument av soft law-karaktär. I fråga om vissa avgränsningar kring definitionen av algoritmisk handel finns t.ex. vägledning i ESMA, Questions and Answers On MiFID II and MiFIR market structures topics, ESMA70-872942901-38, 6 april 2021.

[17] MiFID II, skäl 62.

ellt till andra myndigheter som utövar tillsyn över handelsplatsen i fråga (8 kap. 24 § VpmL), och dokumentera de åtgärder som vidtas för att säkerställa kapacitet, driftsäkerhet och så vidare (8 kap. 23 § 4 st.).[18]

Om värdepappersinstitutet bedriver algoritmisk högfrekvenshandel (HFT) tillkommer ett viktigt ytterligare krav, nämligen att värdepappersinstitutet ska bevara korrekta och kronologiska uppgifter om alla sina lagda order, utförda order och bud på handelsplatser så att Finansinspektionen ska kunna granska dessa vid inspektion.[19] Formulär för sådana rapporter finns fastställda av Kommissionen genom en delegerad förordning.[20] Mot bakgrund av den enorma mängden transaktioner sker sådan dokumentation vanligen också med hjälp av artificiell intelligens, vilket med vedertagen term i finansbranschen brukar kallas RegTech (efter ”regulatory technology”, i det här fallet ett något oegentligt begrepp för vad som egentligen är en compliance technology).

2.3 Potentiella problemområden

Regelverkets komplexitet medför att värdepappersinstitut är hänvisade till specialiserade jurister för att få en god bild av hur de ska bedriva sin verksamhet. De värdepappersinstitut som har en hög benägenhet att lojalt följa regelverket kommer då att acceptera de kostnader som följer av sådan regelefterlevnad, medan mindre samvetsgranna aktörer (som kanske egentligen är målgruppen för reglerna) kan förväntas försöka utnyttja systemets komplexitet för att undvika och fördröja sin egen regelefterlevnad. Här blir det centralt att ha en aktiv och tydlig tillsynsverksamhet och kraftfulla sanktioner. Vi återkommer till sanktionssystemet nedan.

Ett annat problem har att göra med vissa vaga uttryck som används i de materiella reglerna. Vad menas med ”effektiva” system och riskkontroller? Visserligen finns i regelverket en del ytterligare information, men *til syvende og sidst* är detta knappast en formulering som lämpar sig för de binära svarsalternativen ”effektivt” respektive ”ineffektivt”. Snarare måste idealtyperna effektivt – ineffektivt ses som extrempunkter på en skala eller ett spektrum, där avgörandet av vad som är tillräckligt effektivt eller

[18] Ytterligare detaljer finns i Kommissionens delegerade förordning (EU) 2017/589 (not 12).

[19] Om legaldefinitioner av algoritmisk handel resp. algoritmisk högfrekvenshandel, se avsnitt 3.

[20] Kommissionens delegerade förordning (EU) 2017/589 (not 8) bilaga II.

inte tillräckligt effektivt rimligen måste vara avhängigt en mängd faktorer i enskilda fall. Det återstår att se hur myndigheter och domstolar kommer att förhålla sig till begreppet ”effektivt” i det här sammanhanget.

3 De relevanta definitionerna

3.1 Algoritmisk handel

Med begreppet ”algoritmisk handel” avses i lagtexten handel med finansiella instrument där en datoralgoritm automatiskt bestämmer enskilda orderparametrar med begränsat eller inget mänskligt ingripande (1 kap. 5 § 1 p. VpmL). Begreppet ”finansiellt instrument” avser i sin tur överlåtbara värdepapper, penningmarknadsinstrument, andelar i företag för kollektiva investeringar, finansiella derivatinstrument och utsläppsrätter (1 kap. 4 § VpmL), medan begreppen ”datoralgoritm” och ”orderparameter” saknar definition i lagtexten.

Den svenska lagtexten i VpmL ska, till den del den införlivar regler ur MiFID II i svensk rätt, inte tolkas med ledning av svenska förarbeten till VpmL utan i ljuset av det underliggande direktivets (MiFID II:s) ordalydelse och syfte.[21] I direktivet finns viss ytterligare ledning. För begreppet ”orderparameter” ges i direktivet ett antal exempel; ”såsom huruvida ordern ska initieras, tidpunkt, pris och kvantitet för ordern eller hur ordern ska behandlas efter det att den har lagts”.[22] Inte heller direktivet ger emellertid någon ledning avseende den legala definitionen av begreppet ”datoralgoritm”. Det är inte långsökt att anta att lagstiftaren har velat undvika tekniska definitioner av vad som är en datoralgoritm, inte minst eftersom den tekniska utvecklingen på området är snabb och svår att förutse. Vi kan kanske därför låta en rättslig förståelse av begreppet utgå från dess kontext i definitionen av algoritmisk handel, vilket skulle medföra att det rör sig om en datorbaserad teknik för automatisk (dvs. med ”begränsat eller inget mänskligt ingripande”) handel med finansiella instrument i enlighet med vissa förprogrammerade preferenser för när och hur en transaktion ska initieras. Just ledet om begränsat mänskligt ingripande utvecklas i Kommissionens delegerade förordning (EU) 2017/565. Enligt den delegerade förordningens artikel 18 ”ska ett system anses ha inget eller begränsat mänskligt ingripande där, för varje

[21] Mål 14/83 *von Colson och Kamann* EU:C:1984:153 p. 26; mål C-371/02, *Björnekulla Fruktindustrier* EU:C:2004:275 p. 13.

[22] MiFID II art. 4 p. 39.

order- eller budgenerationsprocess eller varje process för att optimera orderutförande, ett automatiserat system fattar beslut på alla stadier om när en order eller bud ska initieras, genereras, 'routas' eller utföras utifrån på förhand fastställda parametrar".

3.2 Teknik för algoritmisk högfrekvenshandel

Kraftfulla algoritmer som kan reagera mycket snabbt och initiera många transaktioner på mycket kort tid är av särskilt intresse för det finansiella systemets stabilitet. Lagtexten innehåller därför också en definition, låt vara en tämligen öppen och flexibel definition, av sådan "teknik för algoritmisk högfrekvenshandel", vilket alltså är en underkategori av algoritmisk handel.[23] Enligt 1 kap. 5 § 27 p. VpmL kännetecknas sådan teknik av

- att infrastrukturen är avsedd att minimera fördröjning (latens) genom samlokalisering, närvärdskap eller elektroniskt höghastighetstillträde,
- att systemet beslutar när en order ska initieras, genereras, styras eller utföras utan mänsklig medverkan för enskilda handelstransaktioner eller enskilda order, och
- en stor mängd intradagsmeddelanden som utgör order, bud eller annulleringar.

Definitionens första del tar inte sikte på den handlande algoritmens egenskaper, utan snarare på den infrastruktur (främst lokaler och hårdvara) som möjliggör högfrekvent handel. Eftersom vi här framför allt intresserar oss för regleringen av själva algoritmerna kommer vi nedan att koncentrera oss på den andra och tredje delen av definitionen.

Andra delen handlar om algoritmens arbetssätt, det vill säga att det återigen är det automatiserade systemet som självständigt fattar beslut om att initiera transaktioner. Detta ligger mycket nära definitionen av algoritmisk handel i 1 kap. 5 § 1 p. VpmL, som redovisats ovan. En skillnad föreligger i att algoritmen som används för högfrekvent handel, när den väl är programmerad och försatt i arbete, ska agera "utan mänsklig medverkan", inte bara "med begränsat eller inget mänskligt ingripande". Detta följer rimligen naturligt av hastigheten i handeln, som vi strax ska se. Det finns också en marginell skillnad i de exemplifierande uppräkningarna, som dock får antas sakna rättslig betydelse.

[23] Kommissionens delegerade förordning (EU) 2017/565 (not 12), skäl 24.

Den sista delen handlar om intradagsmeddelanden. Med ”intradag” avses i handel med finansiella instrument att ett instrument både köps och säljs av samma handlare inom samma handelsdag. Ett ”intradagsmeddelande” ska alltså preliminärt förstås så att en handlande algoritm initierar mer än en order, ett bud eller en annullering med avseende på samma värdepapper på samma handelsdag. Det karakteristiska för högfrekvenshandel är alltså att det ska vara fråga om en ”stor mängd” sådana intradagsmeddelanden. Just begreppet ”stor mängd intradagsmeddelanden” utvecklas i Kommissionens delegerade förordning (EU) 2017/565, artikel 19. I artikel 19.1.b föreligger tyvärr en felöversättning i den svenska språkversionen som gör regeln obegriplig.[24] I det följande förhåller vi oss därför till förordningstexten i övriga språkversioner. En ”stor mängd intradagsmeddelanden” föreligger enligt artikeln när en algoritm skickar order, bud eller annulleringar så snabbt att den i genomsnitt sänder två per sekund med avseende på ett enskilt värdepapper eller i vart fall fyra per sekund med avseende på samtliga värdepapper på samma handelsplats. Artikeln innehåller härutöver ett antal bestämmelser om vilka typer av meddelanden som ska inkluderas eller exkluderas vid beräkningen. Handeln sker alltså övermänskligt snabbt – bara algoritmer kan hänga med.

[24] I den svenska språkversionen står det under artikel 19:

1. En stor mängd intradagsmeddelanden i enlighet med artikel 4.1.40 i direktiv 2014/65/EU ska i genomsnitt bestå av något av följande:

a) Minst två meddelanden per sekund, med avseende på varje enskilt finansiellt instrument som handlas på en handelsplats;

b) minst fyra meddelanden per sekund, med avseende på *varje enskilt finansiellt instrument* som handlas på en handelsplats.

I den engelska språkversionen står det:

1. A high message intraday rate in accordance with Article 4(1)(40) of Directive 2014/65/EU shall consist of the submission on average of any of the following:

(a) at least 2 messages per second with respect to any single financial instrument traded on a trading venue;

(b) at least 4 messages per second with respect to *all financial instruments* traded on a trading venue.

Kursiveringar har tillagts här.

3.3 Potentiella problemområden

Att det föreligger en felöversättning av artikel 19 i den delegerade förordningen främjar naturligtvis inte rättssäkerheten och inte heller EU-rättens allmänna rykte om bristande lagstiftningskvalitet. Förhoppningsvis åtgärdas det, och förhoppningsvis har de flesta som läser artikeln tillräcklig insikt i EU-rätten att de snart går till en annan språkversion för att skapa sig en uppfattning om vad som faktiskt avses i artikeln. Icke desto mindre är sådana misstag naturligtvis oacceptabla.

Ett annat potentiellt problem med de givna definitionerna är att de i flera avseenden är ganska öppna. Å ena sidan kan detta ses som en naturlig konsekvens av att vi har att göra med ett teknikområde i snabb utveckling och av lagstiftarens intresse av att kunna fånga in en stor mängd teknik utan att bli alltför låst av skrivningar som är tidsbundna. Å andra sidan skapar det en oförutsebarhet. Detta gäller i synnerhet vad som ska avses med en ”datoralgoritm” och den öppna uppsättningen indikatorer på vad som är utmärkande för teknik för algoritmisk högfrekvenshandel. Det är uppenbarligen en fråga för rättstillämpningen att till sist avgöra hur definitionerna ska förstås i enskilda fall. Svårigheten med att med precision ringa in sådana tekniker som det nu är fråga om märks också i Kommissionens förslag till förordning om artificiell intelligens, där definitionen av just AI-system sker i ganska allmänna ordalag och med hänvisning till en exemplifierande bilaga till förordningstexten.[25] Frågan har ingen enkel lösning och kanske var detta en god kompromiss. Vi vill ändå påpeka att berörda på finansmarknaden kommer att få leva med viss osäkerhet, till dess det genom praxis har skapats en viss stadga kring vad som avses med de begrepp som, trots allt, är centrala för de regler vi nu diskuterar. Denna osäkerhet kan komma att förvärras genom den successiva introduktionen av ny och mer avancerad AI-baserad teknik för algoritmisk handel.

[25] Kommissionen, Förslag till Europaparlamentets och Rådets förordning om harmoniserade regler för artificiell intelligens (Rättsakt om artificiell intelligens), COM(2021) 206 final, art 3(1).

4 Sanktionssystemet i VpmL

4.1 Administrativa sanktioner och åtgärder

Det finns inte några sanktioner som riktar in sig särskilt på att avskräcka från riskabel högfrekvenshandel. De sanktionsregler vi nu ska gå in på är istället delar av den allmänna sanktionsarsenalen. De kan emellertid komma i fråga vid en överträdelse av reglerna om algoritmisk handel.

Om ett värdepappersinstitut åsidosätter de skyldigheter som följer av VpmL har Finansinspektionen såväl skyldigheter som möjligheter att ingripa mot institutet med en rad åtgärder.[26] Enligt 25 kap. 1 § 1 st. VpmL ska Finansinspektionen ingripa mot ett svenskt värdepappersinstitut som har åsidosatt sina skyldigheter enligt VpmL, andra författningar som reglerar företagets verksamhet, företagets bolagsordning, stadgar eller reglemente eller interna instruktioner som har sin grund i en författning som reglerar företagets verksamhet. Lagstiftarens syfte med regelverket i VpmL är bland annat att ge Finansinspektionen tillgång till ett helt batteri med åtgärder ur vilket inspektionen kan välja det slags ingripande som framstår som det mest ändamålsenliga i det enskilda fallet. Tillsynsverksamheten ska bland annat förhindra att värdepappersinstitut ägnar sig åt risktagande som kan hota dess egen stabilitet och som i förlängningen kan hota det finansiella systemets stabilitet, jfr 25 kap. 2 §.

Finansinspektionen får enligt 25 kap. 3 § VpmL förelägga ett svenskt värdepappersinstitut att upphöra med verksamhet som innefattar handel med finansiella instrument på en marknad, om det med hänsyn till reglerna för eller tillsynen över marknaden framstår som uppenbart olämpligt att institutet bedriver handel där. Om värdepappersinstitutet har åsidosatt sina skyldigheter ska Finansinspektionen också, enligt 25 kap. 1 a och 1 b §§ VpmL, ingripa mot någon som ingår i ett svenskt värdepappersinstituts styrelse eller är dess verkställande direktör, eller ersättare för någon av dem.

I 25 kap. 1 § 2 st. ges Finansinspektionen möjlighet att förelägga ett värdepappersinstitut att inom en viss tid begränsa eller minska riskerna i rörelsen i något avseende, att begränsa eller helt avstå från utdelning eller

[26] MiFID II art 70 stipulerar Sveriges skyldighet, som medlemsland i EU, att tillse att ”administrativa sanktioner och andra åtgärder” finns på plats för att motverka överträdelser av reglerna. Dessa ska enligt samma artikel vara ”effektiva, proportionella och avskräckande”. Direktivet föranledde införande av de skrivningar om sanktioner i VpmL som vi nu ska beröra, se prop. 2014/15:57 s. 31 f.

räntebetalningar, eller att vidta någon annan åtgärd för att komma till rätta med situationen. Finansinspektionen kan också meddela förbud att verkställa vissa beslut, eller göra en anmärkning. Myndigheten ska även ingripa genom att utfärda ett föreläggande om det är sannolikt att ett värdepappersinstitut inom tolv månader inte längre kommer att uppfylla sina skyldigheter enligt VpmL eller andra författningar som reglerar bolagets verksamhet. Om värdepappersinstitutets överträdelse är allvarlig har Finansinspektionen möjlighet att återkalla institutets tillstånd, eller (om det är tillräckligt) meddela en varning (25 kap. 5 §). Om ett värdepappersinstitut har fått till exempel sitt tillstånd återkallat, fått en anmärkning eller en varning får Finansinspektionen dessutom besluta att det ska betala en sanktionsavgift (25 kap. 8 §).

För att utföra sitt tillsynsuppdrag har Finansinspektionen vissa befogenheter att granska institutens verksamheter. Myndigheten har rätt att begära in rapporter och andra upplysningar rörande verksamheten (23 kap. 2–3 och 15 §§). Finansinspektionen får även genomföra en platsundersökning hos det berörda institutet (23 kap. 4 §). Om ett värdepappersinstitut inte i tid lämnar föreskrivna upplysningar om sin verksamhet kan myndigheten besluta om förseningsavgift, högst 100 000 kr (25 kap. 11 §). Vid upprepade fall av försenad eller underlåten rapportering kan det bli aktuellt att meddela en varning eller en anmärkning.[27]

Finansinspektionen kan ingripa inte endast mot värdepappersinstitut med verksamhet i Sverige. Inspektionen får också ingripa mot ett svenskt värdepappersinstitut som har brutit mot tillämpliga regler i ett annat land. Beträffande ett utländskt värdepappersföretag finns enligt 25 kap. 12–15 §§ vissa möjligheter för Finansinspektionen att agera om det utländska institutet inte driver sin rörelse i enlighet med VpmL eller andra tillämpliga regler.

Det finns också möjlighet för Finansinspektionen att återkalla ett värdepappersinstuts tillstånd om en person i ledande ställning, med vilket avses någon som ingår i ett svenskt värdepappersinstituts styrelse eller är dess verkställande direktör, eller ersättare för någon av dem, inte längre uppfyller de krav som ställs på vederbörande enligt VpmL (25 kap. 4 §).[28]

[27] Se SOU 2006:50 författningsbilaga s. 323, och prop. 2006/07:115 s. 509 f., och 642.

[28] Se beträffande fysiska personer i utländska företag, 25 kap. 15a–15c. Angående kraven på styrelse och VD m.fl., se Behrendt Jonsson, B, Not suitable to lead an investment firm: Fit and proper assessments as an instrument of accountability. I: Bergström, C F, och Strand, M, (red.), Legal Accountability in EU Markets for Financial Instruments: The Dual Role of Investment Firms, Oxford University Press 2021, s. 205–225.

Tillståndet kan till exempel återkallas om personen i fråga inte längre har tillräcklig insikt och erfarenhet för att delta i ledningen av ett värdepappersinstitut eller på annat sätt är olämplig för en sådan uppgift (jfr 3 kap. 1 § p. 5). Återkallelse får dock bara ske om Finansinspektionen först beslutat att påtala för värdepappersinstitutet att personen eller personerna inte uppfyller kraven, och denna eller dessa personer trots detta sitter kvar på sin post efter det att en av myndigheten bestämd tid på högst tre månader har gått ut (25 kap. 4 § 1 st.). Syftet med regeln är att få värdepappersinstitutet att se till att kraven på lämplighet uppfylls av samtliga personer i ledningen (jfr 3 kap. 1 § p. 5 och 6). Istället för att återkalla tillståndet får dock Finansinspektionen besluta att en styrelseledamot eller verkställande direktör får lämna sin post. Vanligtvis finns det en suppleant som kan träda in i stället för den person som avgått, men finns det inte det får myndigheten tillförordna en ersättare. Ersättarens uppdrag gäller till dess att institutet utsett en ny styrelseledamot eller verkställande direktör. Enligt förarbetena till VpmL ska detta alternativ väljas istället för återkallelse i de fall då de finns brister i ledningen i ett i övrigt livskraftigt institut som inte förmår ändra situationen. Finansinspektionen är dock inte skyldig att välja detta alternativ istället för att återkalla tillståndet utan får göra en bedömning från fall till fall.[29]

4.2 Val av åtgärd

Som ovan nämnts ska Finansinspektionen, om myndigheten upptäcker brister i ett värdepappersinstituts regelefterlevnad, ingripa genom föreläggande, förbud eller anmärkningar. Om institutets regelöverträdelse är allvarlig ska dock Finansinspektionen återkalla institutets tillstånd eller, om det är tillräckligt, meddela en varning. Vid mindre allvarliga överträdelser kan myndigheten istället välja mellan att förelägga institutet att vidta en åtgärd, förbjuda verkställighet av ett beslut eller meddela en anmärkning. Finansinspektionen har alltså möjlighet att välja – dock inte helt fritt – vilken sanktion som ska följa på en regelöverträdelse från institutets sida.[30] Vid valet av sanktion ska hänsyn tas till hur allvarlig överträdelsen är, och hur länge den pågått. Särskild hänsyn ska tas till överträdelsens art, överträdelsens konkreta och potentiella effekter på det

[29] Se SOU 2006:50 författningsbilaga s. 317–18 samt prop. 2006/07:115 s. 638.

[30] För en kritisk analys av Finansinspektionens praxis kring sitt val av åtgärder, se Larsson Klevhill, M, och Persson, A H, a.a.

finansiella systemet, skador som uppstått och graden av ansvar (25 kap. 2 §). I förarbetena till VpmL anges att vid valet av sanktion bör valet falla på den som är mest verkningsfull i det enskilda fallet. Föreläggande kan dock endast användas när det krävs för att få ett institut att vidta åtgärder för att rätta till något. Det kan handla antingen om att institutet ska göra rättelse eller att institutet ska ombesörja något som det inte tidigare gjort. Anmärkning bör användas när det inte finns något att åtgärda, men överträdelsen bör sanktioneras. Anmärkning och föreläggande bör inte användas vid samma överträdelse.[31]

När det gäller valet mellan återkallelse av tillstånd och varning bör den sistnämnda sanktionen väljas när det i och för sig finns förutsättningar för återkallelse, men det i det särskilda fallet framstår som tillräckligt med en varning. Enligt förarbetena till VpmL kan varning anses vara en tillräcklig åtgärd om institutet inte befaras upprepa överträdelsen och prognosen för institutet därför är god, eller att man från institutets sida inte förstod bättre när överträdelsen skedde. Finansinspektionen avgör när omständigheterna är sådana att tillståndet ska återkallas, men det bör inte ske utan särskilda skäl.[32]

Om en överträdelse är mindre allvarlig kan Finansinspektionen utfärda en anmärkning. Förarbeten till VpmL saknar exempel på situationer där en anmärkning kan vara lämplig som sanktionsåtgärd. Däremot anges att på samma nivå som anmärkning ligger ingripande genom att myndigheten förbjuder verkställighet av beslut respektive förelägger institutet att vidta en åtgärd inom viss tid.[33]

Det finns också, enligt 25 kap. 2 § VpmL, möjlighet för Finansinspektionen att avstå från att ingripa om överträdelsen är ringa eller ursäktlig eller om ”någon annan myndighet eller något annat organ” har vidtagit åtgärder mot värdepappersinstitutet och dessa åtgärder bedöms som tillräckliga även ur FI:s perspektiv.[34] Ett sådant ”annat organ” som avses kan enligt förarbetena vara ”en börs disciplinnämnd”.[35] Vi kommer att återkomma till börsernas eget sanktionssystem nedan, men det ska alltså noteras redan här att förekommande disciplinära beslut som fattats av en

[31] Se SOU 2006:50 del 1 s. 443, SOU 2006:50 författningsbilaga s. 316, och prop. 2006/07:115 s. 636, prop. 2013/14:228 s. 311 och prop. 2014/15:57 s. 68, 72.

[32] Se SOU 2006:50 del 1 s. 443 och prop. 2006/07:115 s. 636 f. Se även prop. 2002/03:139 s. 383.

[33] Jfr prop. 2006/07:115 s. 636 och prop. 2002/03:139 s. 383 och s. 548.

[34] Se vidare prop. 2006/07:115 s. 636 och prop. 2002/03:139 s. 384.

[35] Prop. 2016/17:162 s. 754, med hänvisning till prop. 2006/07:115 s. 637.

börs kan påverka FI:s val av åtgärd på så sätt att Finansinspektionen avstår från att ingripa. Det finns inga andra hänvisningar till ”andra organ” i VpmL.

4.3 Sanktionsavgifter

Om ett värdepappersinstitut har meddelats beslut om anmärkning eller varning av Finansinspektionen får inspektionen som ovan nämnts besluta att institutet ska betala en sanktionsavgift (25 kap. 8 §). Avgiften tillfaller staten. I förarbetena till VpmL sägs att syftet med sanktionen är att den ska tillföra en ekonomiskt omedelbart kännbar sanktion och ge Finansinspektionen möjlighet att gradera en överträdelse. Myndigheten har behörighet att fastställa hur stor avgiften ska vara och behöver således inte vända sig till domstol för att få den utdömd. Enligt 25 kap. 9 § ska sanktionsavgiften fastställas till lägst fem tusen kronor och högst till ett belopp i kronor som motsvarar fem miljoner euro. Avgiften får dock inte överstiga tio procent av institutets omsättning (eller i förekommande fall motsvarande omsättning på koncernnivå) under närmaste föregående räkenskapsår, eller två gånger den vinst som företaget gjort till följd av regelöverträdelsen (om det beloppet går att fastställa). Om överträdelsen har skett under institutets första verksamhetsår eller om uppgifter om omsättningen saknas eller är bristfälliga får omsättningen uppskattas. Avgiften får inte vara så stor att institutet därefter inte uppfyller kraven på soliditet och likviditet (8 kap. 3 §). I 25 kap. 10 § framgår att när sanktionsavgiftens storlek ska fastställas ska särskild hänsyn bland annat tas till hur allvarlig överträdelsen är som föranlett ingripandet och hur länge överträdelsen har pågått. Sanktionsavgiftens storlek ska således ses som ett sätt att ytterligare gradera överträdelsen.

5 Börsernas självreglering och egentillsyn

Ett högt förtroende för börsernas självreglering och egentillsyn är en förutsättning för en väl fungerande värdepappersmarknad. I Sverige finns en lång och stabil tradition av självreglering på finansmarknaderna. Börsbolagen och deras ägare har intresse av att bidra till en god etik på värdepappersmarknaden, och av att övriga aktörer bidrar genom att gemensamt utforma och besluta om regler och god sed. Bolagsstyrningskoden[36] och takeover-regleringen[37] är exempel på områden där Sverige valt självreglering före lagstiftning. En fördel med självregleringen är att finansbranschens centrala aktörer själva utformat reglerna. God förankring och förväntningar på god efterlevnad uppnås därmed bland de som ska tillämpa reglerna. Andra fördelar är flexibilitet, det vill säga förmåga till snabb anpassning till förändrade omständigheter, och med stor sannolikhet lägre kostnader för efterlevnad vid en jämförelse med icke marknadsnära lagstiftning. Självreglering anses också spela en viktig roll för att undvika en alltför detaljerad lagstiftning. Den alltmer tilltagande och detaljerade EU-regleringen, med ett ökat antal tvingande regler, innebär samtidigt utmaningar för den svenska självregleringen och medför krav på att hitta lösningar som kan utgöra alternativ till EU-reglering eller kanske en metod för att genomföra EU-reglering på området.[38] En del områden som traditionellt omfattats av den svenska självregleringen har sedermera övertagits och täckts in av EU-lagstiftning.[39] På andra områden har nya former utvecklats för att nyttja självregleringen inom dessa nya europeiska ramar.[40] En bärande del av självregleringen är att det ställs krav på egen övervakning och tillsyn över aktörerna på marknaden. För det ändamålet har en organisation med nämnder byggts upp. Till exempel har myndighetsuppgifter på värdepappersområdet delegerats av Finansinspektionen till organ som Aktiemarknadsnämnden och Nämnden för svensk redovisningstillsyn, något som markerar egentillsynens starka

[36] Svensk kod för bolagsstyrning, 1 januari 2020.

[37] Genom lagen (2006:451) om offentliga uppköpserbjudanden på aktiemarknaden, genomförs EU:s direktiv 2004/25/EG (”takeoverdirektivet”). För en kommentar och analys, se Stattin, D, Takeover – Offentliga uppköpserbjudanden – Reglering, tolkning och tillämpning, 2006.

[38] Om risken för fragmentering av de nationella regelverken till följd av ”metodlöst” och oreflekterat införlivande av EU-rättsliga regler, se t.ex. Strand, M, EU och civilrättens splittring: Exemplet preskription och ränta vid skadestånd, Tidsskrift for Rettsvitenskap 130(4) (2017) 313.

[39] Ett exempel är det tidigare kravet på licens för försäkringsmäklare, som numera är ett krav på tillstånd från Finansinspektionen för försäkringsförmedlare genom EU-direktivet (2016/97) om försäkringsdistribution.

[40] Som exempel kan nämnas Nämnden för redovisningstillsyn som inrättades 2019 efter förslag av utredningen om ändrade informationskrav på värdepappersmarknaden i slutbetänkandet En ny ordning för redovisningstillsyn (SOU 2015:19).

ställning.[41] Mandaten för nämnderna erkänns på kontraktsrättslig grund genom medlemskap i föreningar eller på marknadsplatser.

För vår undersökning är det särskilt relevant att se närmare på en nämnd med mandat att utdöma sanktioner för att stävja brott mot självregleringen, och som har bäring på algoritmisk handel: Nasdaq Stockholms disciplinnämnd (jfr 13 kap. 14–16 §§ VpmL). Nämnden har till uppgift att pröva ärenden om börsmedlemmars och de noterade bolagens överträdelser av de regler som gäller vid börsen. Om börsen misstänker att en medlem eller ett bolag handlat i strid med regelverket anmäls detta till disciplinnämnden. Börsen utreder och driver ärendet och disciplinnämnden bedömer fallet och fattar beslut om eventuella sanktioner. Sanktioner för ett noterat bolag kan vara varning, vite eller avslutat medlemskap (dvs. avnotering). Vitesbeloppet kan uppgå till mellan 1 och 15 årsavgifter. Exempel från praxis är att det högsta vitesbeloppet utdömdes för Swedbanks bristande aktiviteter mot penningtvätt men av nämndens årsberättelser följer att ett vanligare vitesbelopp är två årsavgifter motsvarande 400 000 kronor (2021). Bland de mer frekventa brister som resulterar i viten ingår bristande transparens och försenad rapportering som kan leda till otillåtna insideraffärer och marknadsmissbruk.

Det följer av 13 kap. 1d § 1 st. VpmL att en börs ska inrätta effektiva system, förfaranden och arrangemang för att säkerställa att deltagare som använder algoritmiska handelssystem på en reglerad marknad som börsen driver inte kan skapa eller bidra till otillbörliga marknadsförhållanden på marknaden och för att kunna hantera eventuella otillbörliga marknadsförhållanden som kan uppstå till följd av användningen av sådana algoritmiska handelssystem. Det följer av andra stycket att det i förfarandena ska ingå krav på deltagarna att utföra lämpliga tester av algoritmer och att tillhandahålla miljöer för att underlätta sådana tester, system för att begränsa andelen inte utförda order i förhållande till transaktionerna som kan läggas in i systemet av en deltagare, system för att det ska vara möjligt att bromsa orderflödet om det finns en risk för att taket för systemkapaciteten uppnås, och system för att begränsa och upprätthålla den minsta prisändring som får tillämpas på den reglerade marknaden.[42] Som

[41] Rätten till delegation följer av 13 kap. 17 § VpmL och genomförs på värdepappersområdet genom Finansinspektionens föreskrifter FFFS 2007:17.

[42] Det ska tilläggas att flertalet börsregler som kan aktualiseras vid algoritmisk handel inte är inriktade just på sådan handel utan är allmängiltiga och teknikneutrala. Om det t.ex. förkommit marknadsmissbruk genom algoritmisk handel så är inte tillämpningen

exempel på hur en marknadsplats reglerar frågan i enlighet med 13 kap. 1d § 1 st. VpmL kan nämnas Nasdaq Nordics regler för sina medlemmar. Här följer att den medlem som vill använda sig av algoritmisk handel ska särskilt garantera att algoritmen inte är skadlig och att den är testad för sitt ändamål, samt anmäla en person med huvudansvar för algoritmen. Sanktioner som kan bli aktuella inkluderar varning, vite och avstängning från marknadsplatsen.[43]

Någon nämnvärd praxis vis-à-vis algoritmisk handel finns inte publicerad. Med en alltmer komplex marknad finns dock anledning att noga följa utvecklingen och ha beredskap för att särskilt utformade regler om sanktioner vid algoritmisk handel behövs. EU:s tillsynsmyndighet på värdepappersområdet, ESMA, tog ett initiativ på området under 2021 och samlade fakta genom en enkät riktad till myndigheter och marknadsaktörer.[44] Enkäten handlade bland annat om ändamålsenligheten av de informations- och sanktionsregler som finns idag på marknaden och behovet av kompletteringar. En slutrapport med tekniska rekommendationer skickades till EU-kommissionen i september 2021.[45] Man kan därför ha hopp om en vidareutveckling av MiFID-reglerna om algoritmisk handel med finansiella instrument.

6 Civilrättsligt ansvar?

Det finns i den EU-rättsliga doktrinen en diskussion kring civilrättsligt ansvar för överträdelser av reglerna om handel med finansiella instrument.[46] Här finns enligt vår uppfattning en tydlig lucka i regelverket. Administrativa sanktioner och straffsanktioner kan vara goda verktyg för att skapa en hög grad av regelefterlevnad i branschen, men de investerare

beroende av om algoritmer använts eller inte utan det handlar mer om att effekten av aktiviteterna på marknaden är av betydelse.

[43] Se vidare Nasdaq Nordic Member Rules, Version 4.0, September 1, 2021, avsnitt 4.11 och 4.13.

[44] Consultation paper MiFIDII/MiFIR Review report on algorithmic trading, ESMA-70-156-2368, Consultation period from 18 December 2020 to 12 March 2021.

[45] MiFID II/MiFIR review report on algorithmic trading, ESMA70-156-4572, 28 Sep 2021.

[46] Se i synnerhet Della Negra, F, MiFID II and Private Law: Enforcing EU Conduct of Business Rules, Hart 2019; samt flera bidrag till Bergström, C F, och Strand, M, (red.), Legal Accountability in EU Markets for Financial Instruments: The Dual Role of Investment Firms, Oxford University Press 2021.

som eventuellt har lidit skada av ett fel i en handelsalgoritm kompenseras inte genom dem. För att investerare ska få kompensation krävs ett system för utkrävande av civilrättsligt ansvar.

Frågan om civilrättsligt ansvar för skada som uppstår vid algoritmisk handel och som drabbar investerare (inklusive konsumenter) är, ur ett svenskt civilrättsligt perspektiv, en inomobligatorisk fråga om felansvar inom ramen för avtal om finansiella tjänster. Det området faller utanför den gällande lagstiftning som finns,[47] och den dispositiva rätt som finns att tillgå består därför av analogier från närliggande lagstiftning och rättspraxis, samt av allmänna kontraktsrättsliga principer.[48] Följaktligen regleras det civilrättsliga ansvaret främst av avtalen mellan värdepappersinstituten och deras kunder. Vi har för avsikt att inom ramen för ett pågående forskningsprojekt[49] bland annat undersöka allokeringen av ansvar inom sådana avtal. För närvarande nöjer vi oss med att konstatera att det föreligger stora oklarheter i det allmänna ansvarssystemet såvitt avser civilrättsligt ansvar.

7 Särskilt om ansvar avseende robotrådgivning

För den småsparare som placerar kapital hos ett värdepappersinstitut är kanske inte den huvudsakliga frågan som inställer sig huruvida kapitalförvaltningen sker genom mänskliga beslut eller genom automatiserad högfrekvenshandel. Småspararen stöter oftare på artificiell intelligens i form av en så kallad rådgivningsrobot. Här finns delvis andra ansvarsfrågor, som vi nu (exkursivt) vill ägna lite utrymme.

Enligt föreningen Sveriges Konsumenter[50] finns det ur ett konsumentperspektiv stora vinster med användningen av AI för finansiella tjänster, så kallad robotrådgivning. Vinsterna för konsumenten består bland annat

[47] Se t.ex. 2 § produktansvarslagen (1992:18). Detsamma gäller på konsumentområdet, vilket följer av 1 § konsumenttjänstlagen (1985:716). Om produktansvar och AI-teknik mer generellt, se Sandra Fribergs bidrag till denna volym.

[48] Se också t.ex. Johan Axhamns och Mikael Hanssons bidrag till denna volym.

[49] Projektet ”AI and the Financial Markets: Accountability and Risk Management with Legal Tools” pågår under perioden 2021–2023. Förutom artikelförfattarna deltar Johanna Chamberlain, Andreas Kotsios och Ensieh Mahi i projektet.

[50] Se Akdag, S, Hur ska konsumenter få rätt mot ett ai-system?, 2020-05-25, http://nyteknik.se/opinion/hur-ska-konsumenter-fa-ratt-mot-ett-ai-system-6995773.

i att denne med några knapptryckningar kan få hjälp med hur vederbörande ska placera sina pengar och få den högsta avkastningen.[51]

En roboträdgivare är enkelt uttryckt en algoritm som t.ex. genererar ett förslag om placeringen av privatpersons sparande så att denna ska nå maximal avkastning med minsta möjliga risk. På den svenska marknaden finns minst fem olika kategorier av robotrådgivare, nämligen fondrobotar, aktierobotar, pensionsrobotar, sparrobotar och bolånerobotar.[52] Effekterna av den digitala rådgivningen har diskuterats i ett flertal forskningsartiklar, i vilka man bland annat har lyft fram att konsumenterna dels har bristfällig kunskap om finansiella instrument, dels bristande kunskap om AI.[53] Resultatet eller rådet som konsumenten får angående hur tillgångar ska placeras kan dessutom skilja sig åt mellan olika system för robotrådgivning. Vidare kan det frågeformulär som roboten utgår ifrån vara ofullständigt vilket gör att den riskprofil som styr robotens förslag till placeringar inte korrekt speglar konsumentens egentliga önskemål och situation.[54]

Med tanke på konsumenternas bristfälliga kunskaper dels om AI, dels om finansiella instrument[55] kan det vara svårt för konsumenten att förstå hur tjänsten fungerar och då tolka dess resultat på rätt sätt. I förhållande till en fysisk, traditionell rådgivare kan en robotrådgivare i regel inte ställa relevanta följdfrågor. Det kan leda till att konsumenten får felaktig in-

[51] Jung, D, Dorner, V, Glaser, F, och Morana, S, Robo-Advisory Digitalization and Automation of Financial Advisory, Business and Information Systems Engineering 60 (1) (2018) s. 81–86; Bachinskiy, A, The growing impact of AI in Financial Services: Six examples, 2019, http://Towardsdatascience.com/the-growing-impact-of-ai-in-financial-services-six-examples-da386c0301b2#:~:text=Predictions%20for%20the%20soon-to-come%20AI%20applications%20in%20financial,as%20the%20adoption%20of%20blockchains%20and%20cryptocurrency%20expands.

[52] Pensionsmyndigheten, Robotrådgivare: En marknadsöversikt, 2018.

[53] Se bl.a. Lewis, D, Computers May Not Make Mistakes but Many Consumers Do. HCI in Business, Government, and Organizations, 2018, s. 361–371 och där angivna källor; och Jung, D, Glaser, F, och Köpplin, W, Robo-Advisory: Opportunities and Risks for the Future of Financial Advisory, 405–427. I: Advances in Consulting Research: Contributions to Management Science, Springer 2019.

[54] Jfr Eriksson, K, Persson, A H och Söderberg, I, Bankrådgivning och rådgivningslagen, KTH 2009, https://docplayer.se/5854644-Bankradgivningsrelationen-och-radgivningslagen.html, där det bl.a. framkommer att uppfattningen om hur riskaversiv som kunden är skiljer sig mellan rådgivare och kund. Undersökningen avser här traditionell rådgivning.

[55] Se Almenberg, J och Säve-Söderberg, J, Financial Literacy and Retirement Planning in Sweden, Journal of Pension Economics and Finance 10(4) 2011 s. 585.

formation, vilket i förlängningen kan leda till att konsumenten väljer en investeringsstrategi som inte är lämplig. Det är också kunden som ska avgöra hur lång sparhorisont som är lämplig och hur hög risknivån ska vara. Roboträdgivaren kan ha svårt att få hela bilden av kundens ekonomi utifrån den information som konsumenten lämnat och föreslår därför kanske ”fel” beslut utifrån kundens perspektiv. I litteraturen om roboträdgivning har allt detta diskuterats, men även tilliten till, transparens inom, förklarbarhet och ansvarsfrågor vid algoritmdrivna processer analyserats.[56] Ytterligare frågor att analysera är t.ex. om det är möjligt att förklara för konsumenten, på ett transparent sätt, hur ett AI-system har fattat beslut, och hur man minimerar risken för att systemet fattar ett ur konsumentens nyttoperspektiv bristfälligt beslut. Organisationen Sveriges konsumenter pekar på att det är mycket svårt för konsumenter att kunna bevisa att de har blivit felaktigt behandlade när bevisningen är dold i en ”algoritmisk djungel”.[57] Det är också svårt för tillsynsmyndigheterna att kartlägga marknaden när priset eller produkten anpassats på en individnivå.

[56] Se Larsson, S, Anneroth, M, Felländer A, Felländer-Tsai, L, Heintz, F, Cederring Ångström, R och Åström, F, Hållbar AI: Inventering av kunskapsläget för etiska, sociala och rättsliga utmaningar med artificiell intelligens, AI Sustainability Center 2019, https://live-aisc.pantheonsite.io/wp-content/uploads/2021/04/Hallbar_AI_web.pdf och angivna källor i rapporten; Norrbin, F och Stenbeck, E, Vägen till lyckad roboträdgivning: En kvalitativ studie om kundförtroende och transaktionskostnader, Linköpings universitet 2018, http://liu.diva-portal.org/smash/get/diva2:1245458/FULLTEXT01.pdf. Vid ett seminarium i Konsumentverkets regi lyftes bl.a. frågan om konsumentskyddet är tillräckligt på digitala och datadriva marknader. Se också Konsumentverket, Seminarium om digitaliseringens effekter hos konsumenter, 23 oktober 2019, https://www.konsumentverket.se/aktuellt/nyheter-och-pressmeddelanden/nyheter/2019/konsumenternas-stallning-pa-digitaliserade-och-datadrivna-marknader/.

[57] Se Akdag, a.a. Möjligtvis kan det bli lättare för konsumenterna i framtiden när det gäller kreditavtal. Se Förslag till Europarlamentets och Rådets direktiv om konsumentkrediter, Bryssel den 30.6.2021 COM(2021) 347 final 2021/0171 (COD), där det framkommer i art. 18 (6) att om kreditprövningen inbegriper användning av profilering eller annan automatiserad behandling av personuppgifter ska medlemsstaterna säkerställa att konsumenten har rätt att (a) begära och erhålla ett mänskligt ingripande från kreditgivaren eller leverantören av tjänster för lånebaserad gräsrotsfinansiering i syfte att få en omprövning av beslutet, (b) begära och erhålla en tydlig förklaring av kreditprövningen från kreditgivaren eller leverantören av tjänster för lånebaserad gräsrotsfinansiering, inbegripet av logiken bakom och riskerna med den automatiserade behandlingen av personuppgifter samt dess betydelse och inverkan på beslutet, (c) uttrycka sin ståndpunkt och bestrida kreditprövningen och beslutet.

Robotrådgivning väcker en mängd rättsliga frågor. Förmår det rättsliga regelverket, till exempel lagen (2003:862) om finansiell rådgivning till konsumenter, tillvarata konsumenternas behov av skydd när rådgivningen sker i en digitaliserad form? Om ett värdepappersinstitut begår överträdelser av regelverket som utmynnar i ett ingripande och en sanktionsavgift är det inget som automatiskt kompenserar kunden för den eventuella förlust som uppkommit på grund av överträdelsen. Utkrävande av civilrättsligt ansvar får göras med utgångpunkt från parternas avtal och tillämplig civilrättslig lagstiftning, men frågan är om det är tillräckligt?

8 Avslutning

Som vi ovan redovisat finns det oro kring algoritmisk handel med finansiella instrument, i synnerhet kring högfrekvenshandel. Vi har ovan citerat skälen till MiFID II, enligt vilka det möjligen kan befaras ”att algoritmiska handelssystem ska överreagera på andra marknadshändelser, vilket kan förvärra volatiliteten om det redan finns ett problem på marknaden”, eller i övrigt ”brista på ett sätt som kan skapa en oordnad marknad”.[58] För att komma tillrätta med detta har det skapats ett regelverk. Vi ser flera styrkor i de regler som antagits i EU och i Sverige, men också flera problem.

Det samlade regelverket är mycket komplext, och består av en väv av EU-akter och svenska författningar. Detta skapar i sig kostnader för regelefterlevnad och kanske också möjligheter att hitta ”kryphål”. Reglerna innehåller även ett flertal vaga uttryck, inte minst i de centrala definitionerna. Det har överlämnats åt myndigheter och domstolar att successivt skapa klarhet kring de här uttryckens närmare tolkning och tillämpning, men till dess så sker råder viss osäkerhet. Det finns också en ren felskrivning i den svenska språkversionen av Kommissionens delegerade förordning 2017/565, vilket är mycket olyckligt. Det civilrättsliga ansvaret är oklart, och kommer huvudsakligen att styras av avtal mellan värdepappersinstitut och investerare (inklusive konsumenter). Här ser vi risker och kanske ett område för kompletterande skyddslagstiftning.

Att det finns vissa problem och luckor i reglerna om algoritmisk handel med finansiella instrument är knappast förvånande. Området är förhållandevis nytt och har de senaste åren varit i snabb och konstant

[58] MiFID II, skäl 62.

utveckling. Dessutom är området, bland annat på grund av de svårigheter med relevanta definitioner som vi ovan redovisat, i sig svårt att reglera. Systemet av materiella regler, tillsynsåtgärder och sanktioner som bygger på självreglering och förtroende mellan parterna är därmed också omoget. Den successiva framväxten av allt mer avancerad AI-teknik kan sannolikt komma att accentuera problemen. Visserligen kan de initiativ som just nu tas kring en allmän reglering av artificiell intelligens på sikt få effekter även för algoritmisk handel med finansiella instrument, men vi kan inte för vår del se att något i Kommissionens förslag till förordning om artificiell intelligens – med dess mycket svaga kopplingar till ansvarsfrågorna – ska få någon större betydelse inom det område vi diskuterat. Vi har större förväntningar på resultaten av ESMA:s pågående initiativ inom ramen för översynen av MiFID II och de förslag till reformer som detta kan leda till.

Med tanke på de stora värden som står på spel för såväl de enskilda konsumenterna som för samhället i stort är det inte några småsaker vi diskuterar. Att följa utvecklingen och successivt bygga ett sammanhållande system för tillsyn och sanktioner – en ansvarets infrastruktur[59] kring algoritmisk handel med finansiella instrument – är en angelägen uppgift för rättssamhället. Vi ser fram emot att få bidra.

[59] Se Strand, M, och Bergström, C F, Investment Firms, Accountability, and the Effective Rule of Law. I: Legal Accountability in EU Markets for Financial Instruments: The Dual Role of Investment Firms, Oxford University Press 2021. Författarna introducerar där begreppet ”infrastructure of accountability” som teoretiskt verktyg för analys av just tillsyns- och ansvarssystem.

Notes on contributing authors

Vladimir Bastidas Venegas. Associate Professor, European Law, Faculty of Law, Uppsala University.

Silvia A. Carretta. Doctoral Candidate, Private Law, Faculty of Law, Uppsala University. Part of the Wallenberg AI, Autonomous Systems and Software Programme – Humanities and Society (WASP-HS) Graduate school.

Anni Carlsson. Doctoral Candidate, Constitutional Law, Faculty of Law, Uppsala University.

Liane Colonna. Assistant Professor, Law and Information Technology, Faculty of Law, Stockholm University. Member of the Swedish Law and Informatics Research Institute (IRI, Stockholm University). Researcher in the Wallenberg AI, Autonomous Systems and Software Programme – Humanities and Society (WASP-HS).

Mattias Dahlberg. Professor, Fiscal Law, Faculty of Law, Uppsala University.

Bruno Debaenst. Senior Lecturer and Associate Professor, Legal History, Faculty of Law, Uppsala University.

Bengt Domeij. Professor, Private Law, Faculty of Law, Uppsala University.

Johan Eddebo. PhD, Philosophy of Religion. Researcher at CRS (Centre for Multidisciplinary Research on Religion and Society), Faculty of Theology, Uppsala University.

Katarina Fast Lappalainen. Assistant Professor, Law and Information Technology, Department of Law, Stockholm University. Member of the Swedish Law and Informatics Research Institute (IRI, Stockholm University).

Stanley Greenstein. Senior Lecturer and Assistant Professor, Law and Information Technology, Faculty of Law, Stockholm University. Member of the Swedish Law and Informatics Research Institute (IRI, Stockholm University).

Mikael Hansson. Senior Lecturer, Private Law, especially Labour Law; Associate Professor, Private Law; Faculty of Law, Uppsala University.

Charlotte Högberg. Doctoral Candidate, Technology and Society, Department of Technology and Society, Lund University. Part of the Wallenberg AI, Autonomous Systems and Software Programme – Humanities and Society (WASP-HS) Graduate school.

Stefan Larsson. Associate Professor, Technology and Social Change, Lund University, Department of Technology and Society. LLM and PhD in Sociology of Law.

Malou Larsson Klevhill. Associate Professor, Law, Luleå University of Technology; Docent, Private Law, Stockholm University; and Senior Lecturer, Business Law, Uppsala University. Researcher in the Wallenberg AI, Autonomous Systems and Software Programme – Humanities and Society (WASP-HS).

Jonas Ledendal. Senior Lecturer, Department of Business Law, Lund University.

Bert Lehrberg. Professor, Civil Law, Faculty of Law, Uppsala University.

Anna-Sara Lind. Professor, Public Law, Faculty of Law, Uppsala University. Scientific leader at the Centre for multidisciplinary studies on Religion and Society, Uppsala University. Member of the Autonomous Systems and Software Programme – Humanities and Society (WASP-HS) management team.

Cecilia Magnusson Sjöberg. Professor, Law & Informatics. Faculty of Law, Stockholm University. Affiliated with the Swedish Law and Informatics Research Institute (IRI, Stockholm University).

Rami Mochaourab. Senior Researcher, Digital Systems Division of RISE Research Institutes of Sweden.

Panagiotis Papapetrou. Professor, Department of Computer and Systems Sciences, Stockholm University; Adjunct Professor, Computer Science Department, Aalto University, Finland.

Annina H. Persson. Professor of private law, KTH, and guest professor, Faculty of Social Sciences, Uppsala University. Researcher in the Wallenberg AI, Autonomous Systems and Software Programme – Humanities and Society (WASP-HS).

Marianne Rødvei Aagaard. Senior Lecturer, Private Law, Faculty of Law, Uppsala University.

Santa Slokenberga. Senior Lecturer, Administrative Law, Faculty of Law, Uppsala University.

Magnus Strand. Associate Professor, European Law; Senior Lecturer, Commercial Law, Uppsala University. Researcher and PI in the Wallenberg AI, Autonomous Systems and Software Programme – Humanities and Society (WASP-HS).

Markku Suksi. Professor, Public Law, Åbo Akademi University, Finland.

Katja de Vries. Assistant Professor, Public Law, Faculty of Law, Uppsala University. Affiliated with the Swedish Law and Informatics Research Institute (IRI, Stockholm University) and the Centre for Law, Science, Technology and Society (LSTS, Brussels).

Annika Waern. Professor, Human-Computer Interaction, Department of Informatics and Media, Uppsala University.

Rebecka Weegar. Assistant Professor, Department of Computer and Systems Sciences, Stockholm University.

Inger Österdahl. Professor, Public International Law, Faculty of Law, Uppsala University.