Responsibility and AI

Council of Europe study
DGI(2019)05

Rapporteur: Karen Yeung

Prepared by the Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT)

1949.2019
COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

DGI(2019)05

A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework

Prepared by the Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT)

Rapporteur: Karen Yeung

French edition:
Responsabilité et IA

Cover design: Documents and Publications Production Department (SPDP), Council of Europe
Photos: Shutterstock

This publication has not been copy-edited by the SPDP Editorial Unit to correct typographical and grammatical errors.

Printed at the Council of Europe

TABLE OF CONTENTS

Introduction

In the terms of reference for the Steering Committee on Media and Information Society (CDMSI) for the biennium 2018 – 2019, the Committee of Ministers of the Council of Europe asked the CDMSI to "study the development and use of new digital technologies and services, including different forms of artificial intelligence, as they may impact peoples' enjoyment of human rights and fundamental freedoms in the digital age, with a view to giving guidance for future standard-setting in this field" and approved the committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT) as a subordinate structure to facilitate the work of the CDMSI.

In its first meeting on 6-7 March 2018, the expert committee decided to focus the study on the the implications of AI decision-making for the concept of responsibility within a human rights framework. Prof. Karen Yeung was appointed as rapporteur for the preparation of the study.

Composition of the Committee of Experts MSI-AUT

Abraham BERNSTEIN, Professor of Informatics, University of Zürich

Jorge CANCIO, International Relations Specialist, Federal Office of Communications, Switzerland

Luciano FLORIDI, Professor of Philosophy and Ethics of Information, Oxford University

Seda GÜRSES, Assistant Professor, Technical University Delft

Gabrielle GUILLEMIN, Senior Legal Officer, ARTICLE 19

Natali HELBERGER, Professor of Information Law, University of Amsterdam

Luukas ILVES (Chair), Deputy Director and Senior Fellow, Lisbon Council

Tanja KERŠEVAN SMOKVINA, State Secretary, Ministry of Culture, Slovenia

Joe MCNAMEE, Independent Consultant

Evgenios NASTOS, Head of Information Unit, Ministry of Digital Policy, Telecoms & Media, Greece

Pierluigi PERRI, Professor of Computer Law, University of Milan

Wolfgang SCHULZ (Vice-Chair), Professor of Law, University of Hamburg

Karen YEUNG, Interdisciplinary Professorial Fellow in Law, Ethics and Informatics, University of Birmingham

Executive Summary

Advanced digital technologies and services, including task-specific artificial intelligence ('AI') bring with them extraordinary promise. They have already generated very substantial benefits, particularly in the form of enhanced efficiency, accuracy, timeliness and convenience across a wide range of digital services.

Yet the emergence of these technologies has also been accompanied by rising public anxiety concerning their potentially damaging effects: for individuals, for vulnerable groups and for society more generally. If these technologies are to be a force for good which enables, rather than undermines, individual and societal flourishing, then it is imperative that we acquire a deeper understanding of these concerns. Not only does this require us to acquire a deeper understanding of their impact on the enjoyment of human rights and fundamental freedoms, but it also entails careful consideration of questions concerning where responsibility should lie for their adverse consequences.

This study begins from the **premise** that, within contemporary constitutional democratic orders, a society's concepts, institutions and practices of responsibility are of critical importance. This is necessary in order to ensure that individuals and organisations are appropriately held to account for the adverse effects of their actions on others, and in order to establish and maintain the foundations for trustworthy and peaceful social cooperation and coordination.

Accordingly, the **purpose** of this study is to examine the implications of advanced digital technologies (including AI) for the concept of responsibility, particularly in so far as they might impede the enjoyment of human rights and fundamental freedoms protected under the ECHR and how responsibility for those risks and consequences should be allocated.

Its **methodological approach** is interdisciplinary, drawing on concepts and academic scholarship from law, the humanities, the social sciences and, to a more limited extent, from computer science. It concludes that, if we are to take human rights seriously in a globally connected digital age, we cannot allow the power of our advanced digital technologies and systems, and those who develop and implement them, to be accrued and exercised without responsibility. Nations bear the primary duty to protect human rights. They must therefore ensure that those who wield and derive benefits from designing, developing and deploying these technologies are held responsible for their adverse impacts. This includes obligations to ensure that there are effective and legitimate institutional mechanisms that will operate to *prevent and forestall* violations to human rights which these technologies may threaten, and to attend to the health of the larger collective and shared *socio-technical environment* in which human rights and the rule of law are anchored. This summary gives a brief overview of the main content of the report.

Chapter 1: Introduction

Chapter 1 outlines **what AI is and how task-specific AI technologies work**. It refers to AI as a set of advanced general-purpose technologies which use techniques from statistics, computer science, and cognitive psychology to enable machines to do highly complex tasks efficiently. These technologies aim either to reproduce or surpass abilities that would require 'intelligence' in humans; e.g. reasoning, autonomy, creativity, etc. It describes how AI technologies work using machine learning, enabling computational systems to learn from

examples, data, and experience and consequently to perform specific tasks intelligently. It explains how machine learning technologies raise issues of responsibility due to their capacity to enable task automation and to enable machines to make decisions and perform tasks to some extent independently from their human developers.

The Chapter draws attention to the capacity of machine learning systems to learn and change over time, dynamically setting their own sub-goals, and their ability to adapt to local conditions via external sensor information or updated input data. Human designers of these systems decide upon and set their initial parameters and the overarching goal which these systems are intended to optimise. At the same time, machine learning systems are designed to operate by making independent decisions that choose between alternatives in ways that are not pre-programmed in advance, and to do so without any human intervention. Because these systems learn dynamically and iteratively from their environment, which is itself often volatile and continuously changing, this has implications for the stability and predictability of their operation. In particular, these systems have the potential to evolve in unexpected ways (section 1.3).

Chapter 1 then explains how, in the context of our contemporary global data infrastructure, AI technologies display a range of other properties that have direct implications for the concept of responsibility, including their:

- inscrutability and opacity
- complex and dynamic nature
- reliance on human input, interaction and discretion
- general purpose nature
- global interconnectivity, scalability and ubiquity
- reliance on large data-sets
- automated, continuous operation, often in real-time
- capacity to generate 'hidden' insight from merging data sets
- ability accurately to imitate human traits
- greater software complexity (include vulnerability to failure and malicious attack)
- capacity to 'personalise' and configure individual choice environments, and
- capacity to configure social choice environments, thus redistributing risks and benefits to optimise a pre-specified goal (section 1.3)

The chapter also explains the **interdisciplinary 'human rights perspective'** adopted in the study, which draws on the human rights and fundamental freedoms protected under the ECHR in order to:

- understand the nature of the risks and adverse consequences generated by advanced digital technologies,
- help identify how responsibility for those threats, risks and consequences should be attributed and allocated, and
- inform consideration of the kinds of institutional mechanisms that may be needed to ensure that human rights are effectively protected.

Finally, the discussion draws attention to existing work concerning the adverse impact of AI technologies on human rights and fundamental freedoms, and upon which the discussion in Chapter 2 seeks to build.

Chapter 2: Threats, risks, harms and wrongs associated with advanced digital technologies

Chapter Two examines a range of adverse consequences potentially associated with the use of advanced digital technologies. It begins by considering the socio-historical context of technological innovation, suggesting that on-going advances in networked digital technologies are likely to prompt far-reaching changes to social and economic life of a scale and magnitude as unsettling and disruptive as the original Industrial Revolution. The resulting 'New' Industrial Revolution now dawning may resemble the original Industrial revolution in that it is likely to generate myriad benefits but, in so doing, might also generate unintended adverse effects that were not recognised at the time of the revolution's unfolding. Accordingly, making reliable predictions about the aggregate, cumulative effects of the current networked digital revolution over time is extremely challenging.

The discussion then considers how the use of algorithmic decision-making ('ADM') systems that rely on data-driven profiling techniques may threaten several human rights (section 2.1), including:

- **rights to a fair trial and to 'due process'** (Art 6 ECHR), particularly where ADM systems are used to automate decisions that significantly affect individuals, yet typically deny the affected individual the opportunity to participate, contest or otherwise challenge the outcome of the decision or the decision-making inputs. Some of these systems are incapable of producing an explanation of its underlying logic in terms that are intelligible and comprehensible to the individual;

- **rights to freedom of expression and information** (Art 10 ECHR), particularly given the powerful influence which global digital platforms now exert over the informational environments of both individuals and societies, in which automated algorithms typically decide how to handle, prioritise, distribute and delete or remove third-party content online, including during political and electoral campaigns. Although platforms have been well-intended in seeking voluntarily to identify and remove 'extremist' content, there are serious risks that these activities may not meet Art 10(2)'s requirements of legality, legitimacy and proportionality for permissible interference with freedom of expression;

- **rights to privacy and data protection** (Art 8 ECHR), due to the reliance of data-driven profiling technologies on the collection and processing of digital data gleaned from tracking the on-line behaviour of individuals at a highly granular level, across a population, the use of these techniques invariably affect the Article 8 right to private and family life. Although contemporary data protection regimes (such as modernised Conv 108) play an important role in safeguarding the rights and interests of data subjects, they might not in practice provide effective and comprehensive protection;

- **rights to protection against discrimination in the exercise of rights and freedoms** (Art 14 ECHR), may be implicated due to the significant risks of bias and discrimination arising from the use of machine learning algorithms, due to the opportunities for bias of the algorithm's developers, bias built into the model upon which the systems are built, biases inherent in the data sets used to train the models, or biases introduced when such systems are implemented in real world settings. Such biases might not only violate the right to protection against discrimination in the exercise of rights and freedoms protected under Art 14, but may also reinforce biases against groups that have historically been

disadvantaged, thereby compounding and exacerbating discrimination and structural disadvantage.

The discussion then considers how data-driven profiling techniques, when employed at scale, may implicate collective values and interests because they make practices of pervasive surveillance, personalisation and manipulation possible at a population level in ways that risk undermining human dignity and autonomy, for example, by systematically treating individuals as objects rather than as moral subjects (section 2.1.2).

The adverse social implications that might accompany the development and use of AI technologies generally, including those that do not rely on the profiling of individuals, are then considered (section 2.2). They include:

- risks of large-scale harm from malicious attacks
- unethical system design or unintended system failure
- loss of authentic, real and meaningful human contact
- the chilling effect of data repurposing
- the exercise of digital power without responsibility
- the hidden privatisation of decisions about public values (including distributive justice) &
- the exploitation of human labour to train algorithms.

Finally, the discussion highlights the power asymmetry between those who develop and employ AI technologies, and those who interact with and are subject to them (section 2.3). While digital service providers (and relevant third parties) that utilise AI systems can acquire very detailed, fine-grained data about the users of their services which they can mine to generate predictions about user traits, tastes and preferences with considerable accuracy, the users themselves (typically) do not understand the complexities of the digital technologies that they use. Nor do they have equivalent access to detailed information about the organisations and firms whose services they use. This opacity and asymmetry not only expands opportunities for potential exploitation, but may substantially threaten collective values and interests that are not readily expressed in existing human rights discourse, including threats to the socio-technical foundations of moral and democratic community. These collective threats and risks are exacerbated by the capacity of these technologies to operate at unprecedented speed and scale, generating novel threats, risks and challenges which contemporary societies have not historically had to confront. At the same time, they are also likely to generate problems of collective action: although the aggregate adverse effects may be very large, the effect on any particular individual may be relatively minor and remedial action may not be sought.

Chapter3: Who bears responsibility for the threats, risks, harms and wrongs posed by advanced digital technologies?

Chapter 3 considers who bears responsibility for the adverse consequences posed by advanced digital technologies. It begins by clarifying what we mean by responsibility and why responsibility matters, emphasising the vital role of our institutions and practices of responsibility in holding to account those whose actions have adverse impacts upon others and on collective interests and values. These institutions and practices serve the vital role of securing and enabling peaceful, trustworthy social co-operation, and giving expression to the rule of law. Although the concept of responsibility can be understood in many different senses, it highlights the distinction between:

- **historic (or retrospective) responsibility:** which looks backwards, seeking to allocate responsibility for conduct and events that occurred in the past; and
- **prospective responsibility:** which establishes obligations and duties associated with roles and tasks that looks to the future, directed towards the production of good outcomes and the prevention of bad outcomes. Prospective responsibilities serve an important guiding function, offering guidance about our rights and obligations vis-à-vis others, and about the way we should behave in our dealings with others.

It argues that we must attend to both the prospective and historic allocation of responsibility for the adverse consequences associated with AI technologies (section 3.2). Only then can we have confidence that efforts will be made to prevent harms and wrongs from occurring (and if they do occur, then these will be brought to an end) as a result of the development and implementation of these technologies. Societies must therefore ensure that they have institutional structures and mechanisms that can be relied upon to ensure appropriate reparation, repair, and the prevention of further harm or wrongdoing arising from the operation of AI technologies.

It then investigates how advanced digital technologies (including AI) implicate existing conceptions of responsibility (section 3.3). To this end, it highlights differences between the concept of moral responsibility, on the one hand, and legal responsibility on the other. Unlike morality, the law has a highly developed system for institutionalising and enforcing responsibility (including the application of sanctions) because it must adjudicate real world disputes. It is also important to bear in mind the distinction between two separate and distinct (albeit sometimes overlapping) types of adverse effect that can arise from the operation of AI systems:

- violations of human rights (including the rights protected under the ECHR) and
- tangible harm to human health, property or the environment.

This study is primarily concerned with analysing responsibility for human rights violations rather than for tangible harm, focusing primarily on those who create, develop, deploy and preside over AI systems and their settings, and on the responsibilities of nation states to ensure that human rights are adequately protected.

Because AI systems can operate in time and space in new and unprecedented ways, these technologies may challenge our existing conceptions of responsibility. Chapter 3 considers two core themes raised in contemporary discussions concerning the adverse effects of AI technologies. First, the role of the tech industry in promulgating and voluntarily committing themselves to abide by so-called 'ethical standards'. It argues that although these voluntary initiatives are in many ways welcome, these codes and standards typically lack any enforcement and sanctioning mechanisms and cannot therefore be relied upon to provide effective protection (section 3.3.1). Secondly, the alleged 'control problem' that is claimed to flow from the capacity of AI-driven systems to operate more or less autonomously from their creators is claimed to create a 'responsibility gap' because the developers of those systems cannot fairly be blamed for their outputs. Chapter 3 demonstrates that the so-called 'control problem' is based on a very particular moral theory of responsibility, one which places undue attention on the conduct of the agent, and fails to give due weight to the interests of victims in security of the person and property (section 3.3.2).

The discussion in Chapter 3 then identifies and briefly outlines a range of **different 'responsibility models'** that could be adopted to govern the allocation and distribution of responsibility for different kinds of adverse impacts arising from the operation of AI systems (section 3.4), including models based on:

- intention/culpability (section 3.4.1)
- risk/negligence (section 3.4.2)
- strict responsibility (section 3.4.3), and
- mandatory insurance schemes (section 3.4.4).

In order to identify which of these models is most suited for the allocation of historic responsibility for the adverse effects of AI systems, the analysis emphasises the importance of distinguishing between human rights violations, on the one hand, and tangible harm to human health, property or the environment on the other (although a single event may result in both tangible harm and a violation of human rights). Responsibility for rights violations of any kind, including human rights violations, is widely understood as 'strict'. Thus, provided that a human rights violation has been established, there is no need for proof of fault. In contrast, the allocation of obligations of repair for tangible harm to health or property may be legally distributed in accordance with a variety of historic responsibility models. Each model strikes a different balance between our interest, as agents, in freedom of action and our interest, as victims, in rights and interests in security of person and property. It is argued that none of these models are self-evidently the 'correct' or 'best' model for allocating and distributing the various threats, risks and harms associated with the operation of advanced digital technologies. Rather, identifying which (if any) of these models is most appropriate will entail a *social policy choice* concerning how they should be appropriately allocated and distributed.

Chapter 3 then draws attention to several acute challenges that arise in seeking to allocate responsibility for the risks and other adverse impacts arising from the operation of complex and interacting socio-technical systems (section 3.5):

a) **the 'many hands' problem**, which arises because the development and operation of AI systems typically entails contributions from multiple individuals, organisations, machine components, software algorithms and human users, often in complex and dynamic environments. The problem of 'many hands' is not new, and rests largely on a 'choice theory' of responsibility in moral philosophy. Contemporary legal systems have developed a relatively sophisticated set of principles and procedures for determining liability involving multiple defendants who can all be understood as having causally contributed to some adverse event. The law's ability to devise practical and effective responses to the many hands problem is partly due to the greater emphasis which the law places on the legitimate interests of victims (and potential victims) in security of the person. In this respect, the law's response differs from choice theories of responsibility in moral philosophy which focus almost exclusively focus on the moral agent. Moreover, in relation to the human rights violations arising from the operation of AI systems, the discussion highlights the importance of mechanisms that prevent and forestall human rights violations arising from the application of advanced digital technologies. The need for effective prevention is particularly important because the aggregate and cumulative effect of these technologies could seriously threaten the collective foundations necessary for

human rights and fundamental freedom to operate in practice. These threats point to the need to enhance and reinvigorate human rights discourse and protection in a data-driven age (section 3.5.1);

b) **Human-computer interaction:** Acute challenges arise in appropriately allocating and distributing responsibility between humans and machines, particularly when there is a 'human in the loop'. A recurring concern has been that, in order to ensure that complex socio-technical systems that incorporate AI always operate in the service of humanity, they should always be designed so that they can be shut down by a human operator. Yet individuals entrusted with the responsibility to supervise the operation of these systems may be understandably reluctant to intervene. This risks turning humans placed in the loop into 'moral crumple zones', largely totemic humans whose central role becomes soaking up fault, although they have only partial control of the system, and who are vulnerable to being scapegoated by tech developers and organisations seeking to avoid responsibility for unintended adverse consequences (section 3.5.2); and

c) **Interacting algorithmic systems:** Even more intractable challenges arise in seeking to identify, anticipate and prevent adverse events that arise from the interactions between complex, algorithm-driven socio-technical systems that can occur at a speed and scale that was simply not possible in a pre-digital, pre-networked age (eg the stock market 'flash crash' of 2010). The unpredictable nature of interactions between multiple algorithmic systems generates novel and potentially catastrophic risks, which we have barely begun to grasp, let alone anticipate and forestall (section 3.5.3).

All of these problems warrant further sustained attention and consideration.

While most of the discussion in Chapter 3 focuses on the responsibility of technology designers, developers and those who own and implement the systems which rely upon these technologies, the discussion in section 3.6 reminds us that it is states that bear the primary obligation to ensure that human rights are effectively protected. It draws attention to the problem of collective action that the operation of AI systems in a global networked age is likely to generate, highlighting the vital importance of a) national legislation to ensure that human rights are protected, b) the need for properly resourced national enforcement authorities with adequate enforcement powers and c) the valuable role which accessible and convenient collective complaints mechanisms, in addition to individual legal remedies, may play to ensure effective human rights protection.

The discussion then draws attention to a range of non-judicial mechanisms that have potential to help secure both prospective and historic responsibility for the adverse impacts of AI systems, including various kinds of impact assessment, auditing techniques and technical protection mechanisms (section 3.7). Technical protection mechanisms, in particular, have considerable promise. This study emphasises the need to embed these mechanisms within a governance framework that enables the relevant technical standards to be set in a transparent and participatory manner, and to ensure independent external oversight and review of their operation.

Before summarising the various findings in Chapter 3, the discussion in section 3.8 briefly considers whether our existing conceptions of human rights, and the mechanisms through which they are protected and enforced, are fit for purpose in a global and connected digital age. It suggests that the power of networked digital technologies that have emerged in recent

years make possible practices and actions that were previously impossible, and thereby create novel threats, risks and forms of wrongdoing. Accordingly, we may need to *reinvigorate human rights discourse* in a networked digital age, in order to protect and nurture the socio-technical foundations necessary for human agency and responsibility, without which human rights and freedoms cannot be practically or meaningfully exercised. The development of an enhanced and reinvigorated conception of human rights could lead to the development of new institutional mechanisms which are better placed to safeguard against the adverse effects of new digital technologies in a data-driven age.

The findings of Chapter 3 are summarised at the end of the chapter (section 3.9).

Chapter 4: Conclusion

Chapter four concludes by summarising the argument made in the preceding sections. It highlights four findings arising from this study:

First, it is vital that we have effective and legitimate mechanisms that will prevent and forestall human rights violations, given the speed and scale at which many advanced digital systems operate in ways that pose substantial threats to human rights without necessarily generating substantial risks of tangible harm. A preventative approach is especially important given that such threats could seriously erode the social foundations necessary for moral and democratic orders, which are essential preconditions for the exercise of individual freedom, autonomy and human rights. This may include both a need to develop collective complaints mechanisms to facilitate effective rights protection, and to enhance and reinvigorate our existing conceptions and understandings of human rights.

Second, the model of legal responsibility that applies to human rights violations is widely understood as one of 'strict responsibility' without the need for proof of fault. In contrast, obligations of repair for tangible harms may be legally allocated and distributed in accordance with a range of responsibility models, each striking a different balance between our interests as agents in freedom of action, and our interest as victims in rights and interests in security of persons and property. Identifying which (if any) of these models is appropriate for preventing the various threats and risks associated with the operation of advanced digital technologies is not self-evident: rather, it will entail a *social policy choice*. In constitutional democratic societies committed to protecting and respecting human rights, states bear a critical responsibility for ensuring that these policy choices are made in a transparent, democratic manner and in ways that will ensure that the policy ultimately adopted will effectively safeguard human rights.

Third, we should nurture and support technical research concerned with securing prospective and historic responsibility for ensuring due respect for many of the values underpinning human rights protection, which may facilitate the development of effective technical protection mechanisms and meaningful 'algorithmic auditing'. This research needs to be developed by interdisciplinary engagement between the technical community and those from law, the humanities and the social sciences, in order to identify more fully how human rights protections can be translated and given expression via technical protection mechanisms embedded within AI systems, and to understand how a human rights approach responds to problems of value-conflict.

Fourth, the effective protection of human rights in a global and connected digital age requires that we have effective and legitimate governance mechanisms, instruments and institutions to

monitor, constrain and oversee the responsible design, development, implementation and operation of our complex socio-technical systems. This requires, at minimum, both democratic participation in the setting of the relevant standards, and the existence of properly resourced, independent authorities equipped with adequate powers systematically to gather information, to investigate non-compliance and to sanction violations, including powers and skills to investigate and verify that these systems do in fact comply with human rights standards and values.

Finally, the study concludes that if we are serious in our commitment to protect and promote human rights in a global and connected digital age, then we cannot allow the power of our advanced digital technologies and systems, and those who develop and implement them, to be accrued and exercised without responsibility. The fundamental principle of reciprocity applies: those who deploy and reap the benefits of these advanced digital technologies (including AI) in the provision of services (from which they derive profit) must be responsible for their adverse consequences. It is therefore of vital importance that states committed to the protection of human rights uphold a commitment to ensure that those who wield digital power (including the power derived from accumulating masses of digital data) are held responsible for their consequences. It follows from the obligation of states to ensure the protection of human rights that they have a duty to ensure that there are governance arrangements and enforcement mechanisms within national law that will ensure that both prospective and historic responsibility for the adverse risks, harms and wrongs arising from the operation of advanced digital technologies are duly allocated.

A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework

by Karen Yeung*

"A great global challenge confronts all those who promote human rights and the rule of law: how can States, companies and civil society ensure that artificial intelligence technologies reinforce and respect, rather than undermine and imperil, human rights?"

David Kaye, Special Rapporteur on the promotion and protection of the rights to freedom of opinion and expression,
United Nations General Assembly (2018)

* With contributions from colleagues Ganna Pogrebna and Andrew Howes, and research assistance from Charlotte Elves and Helen Ryland, The University of Birmingham. I am grateful to Imogen Goold for her advice concerning the content and contours of Anglo-American tort law.

Chapter 1. Introduction

1.1 Scope of this study

This study examines the implications of 'new digital technologies and services, including artificial intelligence' for the concept of responsibility from a human rights perspective. It focuses on technologies referred to as 'Artificial Intelligence' (AI). AI is notoriously difficult to define, and even technical AI researchers do not appear to have settled upon a widely agreed definition. For the purposes of this study, the definition of AI proposed within the EU Commission Communication on AI will be adopted.[1] It provides that:

> 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 (eg voice assistance, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (eg. advanced robots, autonomous cars, drones or Internet of Things applications)…Many AI technologies require data to improve their performance. Once they perform well, they can help improve and automate decision making in the same domain.

Accordingly, this study uses the term AI to describe a set of advanced general purpose technologies that enable machines to do highly complex tasks effectively that draw upon a set of complementary techniques that have developed from statistics, computer science and cognitive psychology.[2] These technologies aim to reproduce or surpass abilities (in computational systems) that would require 'intelligence' if humans were to perform them, including the capacity for learning and adaptation; sensory understanding and interaction; reasoning and planning; optimisation of procedures and parameters; autonomy; creativity; and extracting knowledge and predictions from large diverse digital data.[3] The scope of this inquiry is limited to AI technologies that are currently available (at least as initial research and development demonstrations) or are plausible in the next five years, with a particular focus on technologies leveraging machine learning, and it proceeds on the assumption that advances will continue to improve the performance of task-specific AI rather than the achievement of 'general AI'.[4] It is concerned only with the use of AI as a technology, that is, for the purposes of undertaking useful tasks, rather than as a scientific research tool by academic and other researchers.[5]

It is undeniable that AI technologies have generated extensive benefits, particularly by enhancing the efficiency, accuracy, timeliness and convenience with which many services are provided. Many such applications can be understood as enhancing the practical reach and extending enjoyment of human rights and freedoms. For example, without the use of AI-driven search engines, the massive volume of information now available via the internet would not be practically useful and accessible, thus enhancing the right to freedom of information (protected under Art 10 of the European Convention on the Protection of Human Rights and

---

1 European Commission 2018a. This definition is elaborated more fully in EU High Level Expert Group on Artificial Intelligence 2019b.

2 EPSRC; Hall and Pesenti 2017.

3 EPSRC.

4 Bostrom 2014.

5 The use of machine learning in commercial research and scientific research is not without difficulties. See for example Leonelli 2018; Metcalfe and Crawford 2016.

Fundamental Freedoms, hereinafter 'ECHR'). Many national governments and regional organisations around the world are devoting considerable resources into developing strategies to foster innovation and development in AI technologies based on a widely shared belief that these technologies can and will deliver very significant benefits in terms of enhanced efficiency, productivity and service delivery.[6] Yet early triumphs associated with these advanced networked digital technologies that have fuelled the so-called 'AI boom' and resulting 'AI arms race'[7] have been accompanied by rising public anxiety concerning the potential damaging effects of these technologies for individuals and for society more generally.[8] These concerns have drawn attention to questions about where responsibility lies for these adverse impacts, threats and risks. For this purpose, the importance of responsibility rests on the need to ensure, within constitutional democratic orders, that individuals and organisations are held to account for the adverse 'other-regarding' effects of their actions.[9] Accordingly, the primary purpose of this study is to examine the implications of advanced digital technologies (including AI) for the concept of responsibility, particularly in so far as they might impede the enjoyment of human rights and fundamental freedoms. For this purpose, it considers adverse effects, both intended[10] and unintended,[11] arising from the development and use of AI that can be understood as bearing directly upon the enjoyment of human rights and freedoms. However, the indirect adverse effects of AI, including those associated with the risks of mass unemployment, and other second- or third-order effects are excluded from scope, as are the implications of their use in military applications (including autonomous weapon systems). This is not to suggest that these risks are unimportant, but merely that they raise particular concerns that are beyond the scope of this inquiry.

1.2 Structure of this study

The aim of this study is to examine where responsibility should lie for the adverse individual and societal threats, risks and consequences associated with the actual and anticipated development and application of advanced digital technologies, particularly as they continue to grow in power and sophistication. It adopts what might be understood as a 'human rights perspective', in so far as the human rights and fundamental freedoms protected under the ECHR, can help both to (a) understand the nature of those threats, risks and consequences; (b) help identify how responsibility for those threats, risks and consequences should be attributed and allocated, and (c) consider the kinds of institutional mechanisms that may be needed to ensure that human rights are effectively protected and that responsibility for the protection of human rights is duly allocated.[12] To this end, this study draws on concepts and academic

---

6 The European Commission has committed "at least €20bn" to AI technologies to be spent by 2020 (White 2018) while the UK has recently committed £1bn: UK Department for Digital, Culture, Media and Sport 2018; UK Department for Business, Energy and Industrial Strategy 2018.

7 See Financial Times 2018. On rivalry between China, US, and EU see European Political Strategy Centre 2018.

8 See the literature cited at n.50 below.

9 See section 3.1 for a discussion of the concept of responsibility and its importance. While questions may arise concerning how responsibility for the positive and otherwise beneficial other-regarding effects of AI should be allocated, because the concern of this report is on considering the potential adverse impacts of advanced digital technologies *on human rights*, this study focuses on responsibility for the adverse impacts, threats and risks associated with these technologies.

10 Intentional attacks on others using AI have been described as the 'malicious use' of AI: Brundage et al 2018. Other adverse effects might be intended but not necessarily malicious. See the examples discussed by Sandvig et al 2014.

11 O'Neil 2016.

12 As the Australian Human Rights Commssion has observed, a human rights approach provides 'a more substantive mechanism by which to identify, prevent and mitigate risk' compared to that of 'technology

scholarship from law, the humanities and the social sciences, including moral, legal and political philosophy and political economy, and from computer science, rather than focusing on the case law jurisprudence of the European Court of Human Rights. It proceeds in four chapters.

Chapter 1 provides a basic outline of how these AI technologies work, before identifying the responsibility-relevant attributes or properties which these technologies, and their contemporary and near-term applications, possess.

Chapter 2 examines the potential adverse individual and collective consequences that the application of advanced digital technologies may pose. It begins by focusing on the use of data-driven profiling technologies, highlighting how they may systematically threaten particular rights, as well as threatening more general collective values and interests. It then considers the threats and risks posed by other AI technologies and their contemporary and anticipated applications. Chapter Two concludes by drawing attention to the growing power asymmetry between those with the capacity and resources to develop and employ AI technologies, and the individuals, groups and populations directly affected by their use.

Chapter 3 then considers where responsibility lies for addressing these potential adverse consequences, particularly if they ripen into rights violations and/or harm, including harm to collective values and interests, including those which might threaten the socio-technical foundations of democratic freedom in which human rights are anchored. It considers several legal 'models of responsibility' that might be relied upon to allocate and distribute these risks and consequences. It also identifies several challenges associated with seeking to ascribe and assign responsibility for the operation of highly complex socio-technical systems, which have typically involved multiple organisations, individuals, and interacting software and hardware components. It then identifies a range of potential mechanisms that might help to address some of these challenges in order to secure effective and legitimate human rights protection.

Chapter 4 concludes.

1.3 Understanding the implications of AI for concepts of responsibility

In order to examine the implications of AI for the concept of responsibility from a human rights perspective, it is necessary to acquire a basic understanding of how these technologies are developed and how they operate.

(a) Machine intelligence and machine learning

Much of the excitement about the promise and potential of AI to generate advances and improvements across a wide range of social domains, including industrial productivity, health, medicine, environmental management and food security, rely on the power and potential of machine learning.[13] Machine learning is the technology that allows computers to perform specific tasks intelligently by learning from examples, data and experience.[14] Although

---

ethics' by 'turning concepts of rights and freedoms into effective policies, practices and practical realities. International human rights principles embody these fundamental values, and the human rights approach gives mechanisms and tools to realise them through implementation and accountabilities.' Australian Human Rights Commission 2018:17.

13 Russell and Norvig 2016.

14 Royal Society 2017:16.

machine learning techniques have been available for some time, they have experienced major advances in recent years due to technological developments, enhanced computing power and the radical increase in the availability of digital data. These advances have enabled the development of machines that can now out-perform humans on specific tasks (such as language processing, analysis, translation as well as image recognition) when, only a few years ago, they struggled to achieve accurate results.[15] These technologies are now ubiquitous in the everyday lives of those living in highly industrialised, contemporary societies. In these societies, people now regularly interact with machine learning systems that enable digital services (such as, for example, search engines, product recommendation systems and navigation systems) to provide accurate, efficient responses to user queries in real-time, while continually improving their performance by learning from their mistakes.[16]

(b) Responsibility-relevant properties of AI

In order to identify how advanced digital technologies (including AI) challenge our existing legal, moral and social conceptions of responsibility, it is important to identify the "responsibility-relevant" attributes or properties which these technologies possess, ie. the properties of these technologies that are likely to affect their impact upon others.

Task automation

For this purpose, one of the most important properties of these technologies lies in their capacity to undertake tasks (many of which formerly required human operators) "automatically", that is, without the need for direct human intervention.[17]

Machine autonomy

Advances in machine learning techniques have resulted in the development and increasing use of systems that are not only automated, but they operate in ways that exhibit autonomy. Although the term 'autonomy' is commonly used to describe many AI-enabled applications in public and policy discussion, within the technical community there does not appear to be any widely used consensus about what, precisely, this term means, and the preconditions for characterising a non-human entity as 'autonomous'. However, in the policy literature, the term 'autonomy' is often used to refer to the functional capacity of computational agents to perform tasks independently that require the agent to make 'decisions' about its own behavior without direct input from human operators and without human control. Computational agents of this kind operate by perceiving their environment and adapting their behaviour in response to feedback concerning their own task performance, so that their decisions and actions are thought not to be 'fully deterministic' at the outset (and therefore not fully predictable in advance) due to the almost infinite variety of contexts and environments in which these agents might operate.[18] So understood, autonomy is a range property which may

---

15 Royal Society 2017: 16. For example, radiologists can be outperformed by image recognition algorithms (The Economist 2018a) while lawyers can be outperformed by AI in some of their functions (Mangan 2017).

16 For an example of the co-evolution of human behaviours in response to machine-learning driven navigation systems, see Girardin and Blat 2010.

17 Liu 2016.

18 European Group on Ethics in Science and New Technologies (EGE) 2018. The EGE also observes that there seems to be a push for even higher degrees of automation and 'autonomy' in robotics, AI and mechatronics (a combination of AI and deep learning, data science, sensor technology, IoT, mechanical and electrical engineering) yet at the same time they see 'development toward ever closer interaction between humans and machines' noting that well aligned teams of AI systems and human professionals perform better in some domains than humans or machines separately.

be more or less present in degrees (rather than an all-or-nothing property), depending upon the extent to which human oversight and intervention is required for the operation of the system.[19]

Box 1: Machine autonomy and sensitivity to context

Contrast a self-driving vacuum cleaner with a self-driving car

- Fundamentally the same technical architecture applies: their overarching purpose is set by the system's human designers but both machine agents are capable of determining their own sub-goals in order to achieve that purpose
- The behaviour of both kinds of machine agents cannot be fully determined at the outset
- Each is capable of perceiving their environment and adapting decisions and actions accordingly
- Yet these machines are expected to operate in highly contrasting contexts (home environments are relatively contained and stable in contrast to the dynamism and complexity of on-road conditions)

Accordingly, the greater the stability and predictability of the environment or context in which these systems operate, the more foreseeable their possible outputs and responses. Hence the anticipated behaviour of the self-driving vacuum cleaner is likely to be easier to foresee and anticipate when compared to that of the self-driving car.

Some machine learning systems are distinguished by their capability to learn and change over time, dynamically setting their own sub-goals, and their ability to adapt to local conditions via external sensor information or updated input data.[20] The individual designers of the system may decide and set its initial state and parameters, including the overarching goal that it is intended to optimise, but once deployed, the operation and outputs of the system will evolve with use in different environments. In particular, these computational systems are intended to operate in ways that allow the system to make independent decisions that choose between alternatives in ways that are not pre-programmed in advance, and to do so without any human intervention. Current AI systems cannot determine the overarching goal which the system is designed to optimise (which must be specified by the systems' human developers) but they are capable of determining their own intermediate sub-purposes or goals.

---

19 The range of levels of control or involvement that human operators can have in a system has been described by The Royal Academy of Engineering into four different grades of control: (a) controlled systems: where humans have full or partial control, such as an ordinary car (b) supervised systems: which do what an operator has instructed, such as a programmed lathe or other industrial machinery (c) automatic systems: that carry out fixed functions without the intervention of an operator, such as an elevator, and (d) autonomous systems that are adaptive, learn, and can make 'decisions': Royal Academy of Engineering 2009: 2. The SAE International has developed standard J3016_201806: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems (SAE International 2018) which has been used, for example, by the US Department of Transportation as part of its Federal Automated Vehicles Policy: US Department of Transportation 2017.

20 Michalski et al (2013).

For the purposes of identifying where responsibility lies for the outputs and the consequences of these systems, of particular importance is their *stability and predictability* (see Box 1). Because these systems learn dynamically and iteratively from their environment (which is itself often volatile and continuously changing) this means that these technologies, and their outputs, have the potential to evolve in unexpected ways. This means that, in practice, these technologies are sometimes characterised by their opacity and the unpredictability of their outputs (discussed below), which may have direct implications for whether, and in what ways, the concept of responsibility can be applied to their decisions, actions and the resulting consequences.

In addition to their capacity to operate without direct human oversight and control, these technologies have a number of other responsibility-relevant characteristics, including their:

a. **Inscrutability and opacity:** Concerns about the opacity of these technologies[21] can be understood in three distinct but related senses.[22] First, unlike early forms of AI, including so-called 'expert systems' which relied on rule-based 'if-then' reasoning, contemporary machine learning systems create and utilise more complex models which can make it difficult to trace their underlying logic in order to identify why and how they generated a particular output. While some forms of learning systems enable the underlying logic to be traced and understood (for example, those which utilise decision-trees), others (including those that utilise neural networks and back propagation) do not.[23] Secondly, even for systems that utilise algorithms whose underlying operation and logic can be understood and explained in human terms, those that have been developed by commercial providers may not be openly available for scrutiny because they are the subject of intellectual property rights, entitling the owner of those rights to maintain the secrecy of their algorithms.[24] Thirdly, even if information about a system is provided (such as the technique used to train the machine learning algorithm, or the formal rules of a rule-based computational system), those who lack technical expertise will not be able to understand or meaningfully comprehend this information, effectively reducing the practical transparency of the system.[25] The combined effect of the inscrutability and opacity of algorithms results in their characterisation as 'black boxes',[26] and these properties have direct implications for the transparency, explainability and accountability of and for the applications that utilise them.[27]

b. **Complexity and dynamism**: Technological applications that utilise AI for specific social purposes can be understood as highly complex socio-technical systems, in that both the underlying mechanisms through which they work, and their dynamic and continual interaction with the environments in which they operate, are complex in their operational logic, generating outcomes that are often difficult to predict, particularly for those employing machine learning algorithms.[28] This means that understanding and anticipating

---

21 See Wagner 2017: 36-37.

22 Burrell 2016.

23 A growing body of technical research on 'explainable AI' has emerged, seeking to identify methods through which these systems might be rendered intelligible to humans. This form of opacity recognises human limitations in fully comprehending or explaining the operation of complex systems, because they reason differently to machines: Zalnieruite 2019. See below at section 3.7.1.

24 See for example *State vs Loomis* 881 N.W. 2d 749 (Wis. 2016). Noto La Diega (2018).

25 Burrell 2016: 4.

26 Pasquale 2015.

27 Burrell 2016; Datta et al 2016. Weller 2017; Yeung and Weller 2019.

28 Schut and Wooldridge 2000.

how they function in real world contexts can be extremely challenging, even for those with the relevant technical expertise, typically requiring expertise from multiple domains.

c. **Human input, interaction and discretion:** Although advances in AI are strongly associated with the so-called 'rise of the machines,' it is important to recognise that humans are involved at every stage of the development and implementation of AI-driven technologies: from the origination of ideas and proposal for development, design, modelling, data-gathering and analysis, testing, implementation, operation and evaluation.[29] In addition, these systems are also often expected to operate in real world environments in which the systems are designed dynamically to interact with humans, and in many cases are intended to do so at scale (eg Facebook's News Feed system). In particular, many applications that utilise AI are designed formally to preserve human discretion, so that the system's output is offered to the user as a 'recommendation' rather than executing some pre-specified automated decision or function.[30] Thus, for example, digital product recommendation engines offer product suggestions to users, but the human user retains formal decision-making authority in deciding whether or not to act upon the recommendation and this may have significant implications for the concept of responsibility.[31]

d. **A general purpose technology:** AI technologies can be understood as 'general purpose', in that they can conceivably be applied to an almost limitless range of social domains. This versatility means that AI technologies can be characterised as classic 'dual use' technologies, in that the motivations for their application may range from benevolent, to self-interested through to malevolent.[32]

e. **Global interconnectivity, ubiquity and scalability:** It is important to recognise that the global interconnectivity and reach of the internet (and internet-connected technologies) have enabled the swift roll-out of AI technologies on a massive scale, particularly with the rapid and widespread take-up of 'smart' networked devices. Many AI applications used daily by individuals in the industrialised world have now become ubiquitous. Given the efficiency and convenience which they offer in managing the routine tasks involved in contemporary life, this means that, in practice, it is rapidly becoming impossible to conceive of modern living without them.[33] Yet the reach and penetration of networked data infrastructure, and the take-up of smart connected devices into the global south, remains poor and limited compared with the global north, so that those living in these areas do not have equivalent access to the services and improvements in efficiency and convenience that are available to those living in wealthier, highly industrialised states.[34]

f. **Real-time automated and continuous operation:** The efficiency and convenience which many AI applications offer can be attributed, in no small measure, to their ability to operate automatically and in real-time.[35] Thus, for example, AI-enabled navigation systems can offer invaluable guidance to individuals as they seek to find their way to a destination which is entirely foreign to them by providing real-time guidance concerning

---

29 Bryson and Theodorou 2018.

30 Su and Taghi 2009.

31 See discussion of 'humans in the loop' at section 3.5.2 below.

32 On the self-interested design of algorithmic systems, see the discussion of the SABRE airline reservation system in Sandvig et al 2014. On malevolent applications of AI, see Brundage et al (2018).

33 Zuboff 2015; Royal Society 2017.

34 McSherry 2018.

35 For examples of real-time AI applications, see Narula (2018).

which direction to take and, at the same time, can advise on the anticipated journey time of alternative route options.[36] These applications are possible because of the capacity of AI technologies to collect digital data from sensors embedded into and collected from internet-enabled devices, enabling them to track the activities and movements of individuals at a highly granular level, and often without the individual's awareness. These technological capacities have direct implications for concepts of responsibility in ways that may affect the enjoyment of human rights and freedoms in at least three ways. Firstly, the networked nature of many of these technologies which the internet (and internet connected technologies) have made possible means that they can operate at scale and in real time. As a result, there may be considerable distance in both time and space between the design and implementation of these systems, and the point at which their decisions and consequences arise and are directly and immediately felt. Secondly, this capacity to operate in real-time and at scale generates very significant challenges for their supervision and oversight, discussed more fully below. Thirdly, in order to provide highly personalised advice that is contextualised against wider population trends (eg traffic congestion) in real-time, this necessitates the continuous surveillance of individuals at a population wide-level, entailing constant personal data collection and processing, which necessarily implicates the human rights to, and collective value of, privacy and data protection.[37]

g. **Reliance on large data-sets**: While the model upon which a computational algorithm is based will determine its operation, machine learning systems rely critically on the underlying data-sets for their accuracy and operation.[38] Without access to relevant data sets, machine learning algorithms are but hollow shells. Accordingly, the availability, size, and quality of the underlying datasets upon which algorithms are trained, tested and validated plays a critical role in their performance and the accuracy and legitimacy of their outputs, as does the availability and quality of the data which these systems rely upon during their operation.

h. **Capacity to generate insight from merging data sets:** Much of the excitement surrounding AI technologies arises from their capacity to generate new insight from merged datasets which can then be used to predict and inform decision-making. In particular, a data set might contain fairly mundane, innocuous data about individuals. But when multiple such data sets are merged and mined, this may generate insight that can enable quite intimate personal information to be inferred at a very high level of accuracy.[39] Accordingly, issues concerning how to govern the collection and processing of digital data have far-reaching implications for human rights and for the concept of responsibility which, given the ease and almost negligible cost associated with transferring and copying digital data and the complexity of the contemporary global data eco-system, have become especially challenging and important.

i. **Capacity to imitate human traits**: In recent years, the ability of AI technologies to imitate human traits, including voice simulation, visual representations of human behaviour and robots capable of interacting with humans with apparent emotional sensitivity, has become so high quality that it may be extremely difficult for ordinary humans to detect that those traits are artificially generated. This has provoked concern about their capacity

---

36 Swan 2015.

37 See discussion at Section 2 below.

38 Kitchin 2014. Prainsack 2019.

39 Kosinski et al 2013.

to deceive humans (particularly the production of so-called 'deep fakes') and harnessed for unethical or other malicious purposes.[40]

j. **Greater software complexity:** Machine learning and deep learning systems become progressively complex, not only due to the availability of data, but also due to increased programming complexity. As a result, these systems are subject to three types of vulnerability: first, increased programming complexity increases the propensity of these systems to generate stochastic components (i.e. make mistakes)[41]; secondly, this complexity opens the door to a wide range of adversarial attacks;[42] and thirdly, the unpredictability of their outputs can generate unintended yet highly consequential adverse third party effects ('externalities').

k. **Capacity to 'personalise' and configure individual choice environments:** One way in which AI systems have contributed to the achievement of greater efficiency and precision across a wide range of processes and operations has been through the 'personalisation' of service provision. For example, the use of profiling techniques enables digital retailers (such as Amazon) to provide 'personalised' product recommendations to each customer, based on data-driven predictions (gleaned from the continuous collection and analysis of that customer's digital traces when analysed in conjunction with those of other customers).[43] While the personalisation of digital services and offers benefits to users in reducing the volume of irrelevant offers and services which they receive, it has the effect of segmenting individual users from each other, such that one user sees only his or her personalised informational environment, which may be very different from that seen by other users. When AI driven personalisation takes place routinely and at scale, this risks fostering social fragmentation[44] and eroding social cohesion and solidarity.[45]

l. **Capacity to redistribute risks, benefits and burdens among and between individuals and groups via the use of AI-driven optimisation systems which reconfigure social environments and choice architectures:** AI systems can operate in real time and at a scale via the internet's global networked architecture. As a result, these systems can be configured to operate in a manner designed to optimise the over-arching goal prespecified by its human developers at a scale that was previously impossible in a pre-internet enabled age.[46] The capacity to harness AI systems to personalise the informational choice environments of each individual user is particularly powerful when configured to operate at scale. It enables the design and deployment of AI content-distribution systems aimed at influencing and directing the behaviour of an entire population of users, rather than one isolated user, in accordance with the developer's chosen optimisation function, these systems inevitably prioritise certain values over others, and will do so in ways that configure and shape social and informational environments that may be beneficial for some individuals and groups, while detrimental to others. For example, the optimisation

---

40 The Economist 2017. Chesney and Citron 2019. See discussion at Section 2.2.2 below.

41 Recent research in image recognition demonstrated the lack of ability of technology to distinguish noisy informational inputs: chihuahua dogs pictures were mixed with muffin pictures and the AI algorithm could not tell them apart: Yao 2017.

42 Current AI technologies can be easily and successfully attacked by cybercriminals who can use AI system vulnerabilities for their own benefit. Cybercriminals can falsify voice recognition and CAPTCHA systems to break into personal and business accounts: Polyakov 2018.

43 Yeung 2016.

44 Pariser 2012.

45 Yeung 2018a.

46 Yeung 2016.

function of AI-driven navigation systems might be to enable each user to find the fastest possible route to her desired destination, given the volume and location of traffic prevailing at the user's time of travel. The routes identified by the system and recommended to users will, when aggregated, have distributional effects: residents in areas in which traffic is routed being confronted with greater noise levels, vehicle emissions and congestion, while these effects will not be experienced by residents in areas where traffic is not routed. Accordingly, these optimisation systems raise questions about accountability and responsibility for their resulting distributional outcomes, particularly given that there is typically no consultation input or deliberation from affected individuals, groups and populations concerning the distribution of risks and benefits arising from their operation.[47]

m. **The capacity to generate problems of collective action:** The capacity of AI optimisation systems to operate in a highly targeted manner which is personalised to individual users, and to do so at scale across an entire population of users, means that these systems can operate in ways that may have a relatively minor effect at the individual level, whilst having a serious and significant impact at the collective and/or societal level. It is not difficult to imagine circumstances in which the operation of AI optimisation systems may therefore generate a 'collective action' problem. Collective action problems arise when all individuals would be better off cooperating with other users, but each individual user fails to take action because the impact on each individual is too small to justify the effort and resources associated with so doing.[48] Consider, for example, the problem of political micro-targeting and the provision of misleading, inaccurate or dubious political information to individual voters with the intention of encouraging them to vote for a particular candidate. Even if a particular individual is misled into voting for a candidate that she might not otherwise have supported, she is in practice unlikely to be sufficiently motivated to initiate a complaint or other legal proceedings against those responsible for its dissemination. Yet because these effect are felt at the population/collective level, they may pose real and potentially serious threats to the integrity of democratic elections, and to democratic processes more generally.[49] In other words, one of the distinctive and novel challenges which AI systems now pose arises from their capacity to operate in a highly targeted and personalised manner, yet in real-time and at a population-wide scale, which could pose serious societal threats but for which the motivation for any individual to try and counter these threats may be extremely weak.

1.4 Implications for the concept of responsibility from a human rights perspective

The importance of understanding the human rights dimensions of AI is reflected in the various inquiries and reports commissioned and produced by a growing number of civil society organisations and is increasingly the focus of academic scholarship concerned with the 'ethics of AI'[50]. This includes the work of the Council of Europe, including its study on the human

---

47 Yeung 2017a.

48 Olsen 1965

49 UK Information Commissioner's Office 2018. UK House of Commons Digital Culture Media and Sports Committee 2019.

50 See for example Amnesty International 2017; Access Now 2018; Australian Human Rights Commission 2018; Cath 2017; Hildebrandt 2015; Executive Office of the President 2016; The Montreal Declaration for Responsible AI 2017; The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems 2018; Latonero 2019; Mantelero 2018; Raso et al 2018; Risse 2018; Rouvroy 2016; UN General Assembly 2018; Mantalero 2019; Nuffield Foundation and Leverhulme Centre for the Future of Intelligence 2019; EU High Level Expert Group on AI 2019a.

rights dimensions of automated data processing techniques and possible regulatory implications, prepared by the Committee of Experts on internet intermediaries (MSI-NET) (hereafter the 'Wagner Study').[51] The Wagner Study identifies examples of algorithmic decision-making systems currently in use that may violate or undermine the enjoyment of 'the most obviously implicated rights that are to a stronger or lesser degree already in public discussion',[52] including rights to:

- a fair trial and due process (Art 6)[53];
- privacy and data protection (Art 8)[54];
- freedom of expression (Art 10);
- freedom of association (Art 11)[55];
- an effective remedy (Art 13)[56]
- the prohibition on discrimination (Art 14)[57] and
- the right to free elections (Art 3, Protocol 1)[58]

Yet, as the Report commissioned by the Parliamentary Assembly of the Council of Europe (PACE) undertaken by the Rathenau Instituut concluded:

> Despite [the]…wide-ranging impact of digital technologies on human rights, so far little attention has been paid to this crucial topic and there has been scarcely any fundamental political and public debate on it. As a result, a serious erosion of human rights is taking place. Therefore, the human rights debate, which is seriously lagging behind the fast-growing technological developments, needs to be strengthened rapidly.[59]

The present study builds on the Wagner Study by critically examining how advanced digital technologies may implicate the concept of responsibility. Chapter 2 begins by identifying and examining the adverse individual and societal risks posed by AI. It adopts a 'human rights perspective' by focusing on how these technologies may undermine the practical capacity to

---

51 The Wagner Study focused primarily on the implications for human rights of algorithmic decision-making systems that affect the public at large, identifying various human rights concerns triggered by the increasing role of algorithms in decision-making, observing that these concerns are bound to expand and grow as algorithms, automated data processing techniques and related systems become increasingly complex and interact in ways that become 'progressively impenetrable to the human mind': Wagner Study 2017: 5.

52 Wagner Study 2017: 32.

53 See Section 2.1.1(a).

54 See Section 2.1.1(b).

55 Although the internet and social networking sights have enhanced the capacity for individuals to exercise their Art 11 ECHR rights to freedom of association, there are concerns that the automated sorting and profiling of protested on-line may erode these rights: Wagner Study 2017: 23-24.

56 Art 13 ECHR requires that states ensure that individuals have access to judicial or other procedures that can impartially decide on their claims concerning violations of human rights, including on-line violations, including effective non-judicial mechanisms, and to ensure that private sector actors respect those rights by establishing effective complaint mechanisms that promptly remedy the grievances of individuals. Yet the opacity of automated decision-making processes may impede the ability of individuals to obtain an effective remedy and the increasing use of automated decision-mechanisms for complaints handling raises 'serious concerns' about whether such mechanisms can be regarded as offering an effective remedy: Wagner Study 2017: 24.

57 See Section 2.1.1 (d).

58 Art 3 of Protocol 1 ECHR requires states to support the individual right to free expression by holding free elections at reasonable intervals. These elections must enable you to vote in secret. However the rise of social media and the use of automated content recommendation systems may be used for the purposes of political manipulation and could threaten the right to free elections: Wagner Study 2017: 30-32.

59 Van Est and Gerritsen 2017: 46.

exercise particular human rights and freedoms on a *systematic* basis in an era pervaded by advanced AI technologies, rather than engaging in detailed analysis of particular AI applications that may adversely impact specific human rights and fundamental freedoms. Two dimensions of these systematic impacts are considered: firstly, the threats to a set of rights posed by algorithmic decision-making systems.[60] Secondly, the wider adverse collective social impacts of AI technologies (including but not limited to those incorporated into algorithmic decision-making systems), only some of which can be readily expressed in the language of existing human rights discourse. Over time, these wider adverse effects could systematically threaten the socio-technical foundations which the very notion of human rights presupposes and in which they are rooted.

[60] A number of these rights are examined in the Wagner Study 2017.

Chapter 2. Threats, risks, harms and wrongs associated with advanced digital technologies

Many commentators claim that advances in networked digital technologies, including those currently referred to as AI technologies, are powering the emergence of a 'New Industrial Revolution' that will provoke far-reaching changes across every aspect of social life, of a magnitude and scale that will be as disruptive and unsettling as those wrought by the original Industrial Revolution.[61] Before examining the potential threats and risks associated with these emerging technologies, it is helpful briefly to highlight the broader social-political and economic context which affect and condition their development, implementation and adoption, and the broader historic context and experience of modern scientific and technological innovation.

To this end, there may be parallels between the larger societal effects of the original industrial revolution and the anticipated effects of the 'New' Industrial Revolution that is now dawning. For example, while the 19th century Industrial Revolution brought about myriad benefits to both individuals and society, and can be credited with very substantial and widespread improvements to living standards and individual and collective well-being, it generated unintended adverse effects. These include both direct adverse effects on human health and safety associated with early forms of industrial production, and the burning of fossil fuels to power industrial activity which has led to a serious climate change problem at a global scale, and which we have not yet adequately addressed or resolved. Yet the adverse effects on climate change arising from the technologies that provoked the original Industrial Revolution did not become apparent until over a century later, by which time it was too late to address and reverse them effectively. Contemporary societies might now face a similar dilemma. One of the difficulties in seeking to identify and anticipate the larger adverse societal effects of technological innovation arises not only from difficulties in predicting their likely applications and take up, but especially from difficulties in anticipating their aggregate, cumulative effects over time and space.

2.1 The rise of algorithmic decision-making (ADM) systems

Computational systems that utilise machine learning algorithms, combined with the rapid and widespread take-up of 'smart' devices, have fuelled the emergence of algorithmic decision-making systems which seek to harness (and frequently to monetise) the digital data which can now be gleaned by systematically tracking and collecting the digital traces left from individuals' on-line behaviours, and utilising advanced digital technologies (including AI) in order to produce new knowledge that can be used to inform real-world decisions. Many of these systems rely upon data-driven profiling techniques that entail the systematic and bulk collection of data from a population of individuals in order to identify patterns and thereby predict preferences, interests and behaviours of individuals and groups, often with very high degrees of accuracy. These data profiles can then be used to sort individuals to identify 'candidates of interest' with the aim of producing 'actionable insight' – that is, insight that can be used to inform and automate decision-making about individuals by those undertaking the profiling (or their clients).[62] These systems are widely used by retailers seeking to target products to individuals identified as most profitable and most likely to be interested in them,[63]

---

61 boyd and Crawford 2013. Skilton and Hovsepian 2017.

62 Mayer-Schonenberg and Cukier 2013.

63 Draper and Turrow 2017; Gandy 1993.

by political actors and organisations seeking to tailor and target campaign messages to individuals who are identified as most likely to be persuaded by them,[64] and, increasingly, by criminal justice authorities who seek to assess the 'risk' which particular individuals are algorithmically identified as posing to public safety in order to make custody decisions about individuals (whether criminal suspects or those convicted of criminal offences).[65]

It is in this socio-economic context that public anxieties have emerged concerning the *societal effects* of advanced digital technologies (including AI), particularly given the increasing use of data-driven profiling. Recent attention has focused on the way in which social media and other content distribution platforms utilize profiling technologies in ways that have profound implications for the Article 10 right to freedom of expression and information, particularly following the Cambridge Analytica scandal in which it is alleged that millions of profiles of Facebook users were illegally collected to microtarget individuals with political messages with the aim of swaying voter behavior.[66] The following discussion, however, is concerned primarily with the way in which data-driven algorithmic decision-making systems more generally may systematically threaten particular human rights, rather than focusing on their application to specific domains of activity.

2.1.1 How do ADM systems systematically threaten particular rights?

The use of algorithmic decision-making systems may systematically threaten several rights including:

(a) The right to a fair trial and rights of 'due process': **Art 6**.

Many ADM systems utilise data driven profiling techniques to create digital profiles of individuals and groups across a wide range of contexts, sifting and sorting individuals into categories in order to assist decision-making. When used to automate and inform decision-making that substantially affect the rights and significant interests of individuals, data-driven profiling may have serious consequences. For the affected individual, the opportunity to participate in, contest or otherwise challenge the outcome of the decision and/or the underlying reasoning upon which that decision was based, or the quality or integrity of the data that was used to inform the decision, are in practice, almost non-existent.[67] While the right to a fair hearing (per Article 6) encompasses a series of more specific procedural rights,[68] these include a person's right to know the reasons for decisions which adversely and significantly affect that individual, yet the ADM systems used to inform decision-making may not be configured to, nor capable of, produce meaningful explanations in terms that are intelligible to the affected individual, or even (in the case of neural networks that rely on back propagation) in terms that are intelligible to the algorithm developers.[69] These concerns are exacerbated by the opacity of these systems which can arise from their technical complexity, difficulties in assessing the quality and provenance of the underlying training data that was used to train the decision-making model,[70] or because the algorithm enjoys intellectual property protection as a trade secret and therefore need not be publicly disclosed,[71] a stance

64 Gorton 2016.
65 Oswald et al 2018; Ferguson 2016.
66 UK House of Commons, Digital Culture Media and Sport 2019.
67 Hildebrandt 2015; Hildebrandt and Gutwirth 2008.
68 Galligan 1997.
69 Weller 2017; Matthias 2004; Burrell 2016.
70 Lohr et al 2019.
71 Pasquale 2015.

which organisations utilising these systems typically defend on the basis of that it prevents users from 'gaming' the system.[72] Accordingly, these systems risk interfering with rights to due process protected under Article 6 (including the presumption of innocence), particularly in circumstances where the consequences for the affected individual are serious and life-limiting.[73] Particularly worrying is the increasing use of AI systems in criminal justice contexts to inform custodial and sentencing decisions, primarily in the USA, although they are being taken up elsewhere (including the UK).[74] Yet, as Hildebrandt has observed, we have become resistant to the notion that the outcomes of an AI tool might be incorrect, incomplete or even irrelevant with regard to potential suspects.[75]

(b) The right to freedom of expression: Art 10

The operation of algorithmic profiling may significantly affect the Art 10 right to freedom of expression, which includes the right to receive and impart information, given the powerful influence which global digital platforms now exert over our informational environment at both an individual and societal level. For example, automated search engines act as crucial gatekeepers for human beings who wish to seek, receive or impart information, as content which is not indexed or ranked highly is less likely to reach a large audience or to be seen at all. Yet search algorithms are intentionally designed to serve their owner's commercial interests, and are therefore inevitably biased towards certain types of content or content providers. It is typically automated algorithms, rather than humans, that decide how to handle, prioritise, distribute and delete third-party content on online platforms, including content handling during political and electoral campaigns. These practices not only implicate the individual right to freedom of expression, but also Article 10's inherent aim of creating an enabling environment for pluralist public debate that is equally accessible and inclusive to all.[76]

In addition, online platforms are increasingly under pressure to actively counter online hate speech through automated techniques that detect and delete illegal content, particularly following the live video streaming via social media platforms of the attack on civilians by a lone terrorist in Christchurch in early 2019. Article 10.2 provides that any interferences with free expression, which would therefore include algorithmic systems that block access to content through filtering or removal, must be prescribed by law, pursue a specified legitimate purpose outlined in Art 10.2, and necessary in a democratic society.[77] Accordingly, the widespread use of algorithms for content filtering and content removal processes, including on social media platforms also raises rule of law concerns, raising questions of legality, legitimacy and

---

72 Bennett-Moses and de Koker 2017.

73 Davidow 2016.

74 These applications not only implicate the rights under Article 6, but also the Article 5 right to liberty and security of the person, and the non-discrimination principle protected by Article 14.

75 Hildebrandt 2016. Hildebrandt argues that the Art 6 'equality of arms' principle should be re-invented the moment that the public prosecutor, judge or lawyer is unable to check on how the police's AI agent reached its conclusions, and that these AI agents should be required to log their activity and outputs, purposes, and how they reached the outcome to enable proper review . The Rathenau Institut endorses Hildebrandt's views, and has suggested that the Council of Europe consider establishing a framework of minimum norms to be taken into account when a 'court' (interpreted for this purpose as including all decision-making authorities within the legal system, particularly those involved in making custody decisions concerning individuals within the criminal justice system) uses AI – helping to prevent member states from devising their own individual frameworks which is likely to result in uneven and varying degrees of protection under Art 6 ECHR provided by individual member states: Van Est and Gerritsen (2017) 42-43.

76 See UN General Assembly 2018.

77 In line with the jurisprudence of the European Court of Human Rights, any restriction of the freedom of expression must correspond to a 'pressing social need' and be proportionate to the legitimate aim(s) pursued. See *Yildirim v. Turkey*, 18 March 2013, No 3111/10.

proportionality, particularly given that that online platforms often face an unclear legislative framework that encourages them to remove content voluntarily, without a clear legal basis. While their intentions are welcome, there is a lack of transparency and accountability concerning the process or about the criteria adopted to establish which content is 'extremist' or 'clearly illegal'.[78] These arrangements create the risk of excessive interference with the right to freedom of expression, and can be understood as 'handing off' law enforcement responsibilities from states to private enterprises. National legal regimes which require digital intermediaries to restrict access to content based on vague notions such as 'extremism' oblige them to monitor all on-line communication in order to detect illegal content, thereby violating the established principle that intermediaries should not be obliged to conduct general monitoring because of their potential 'chilling effects' on freedom of expression.[79] In addition, process related concerns arise due to the capacity of platforms to decide for themselves what constitutes 'extremist' content and therefore subject to removal: the tools and measures through which identification and removal decisions are made effectively rest with private providers and, unless those measures are not subject to meaningful and effective state oversight, risk exceeding legally and constitutionally prescribed boundaries, thereby contravening the rule of law.[80]

While the imperative of acting decisively against the spread of hate messages and the incitement to racially-motivated offences is indisputable, such practices raise considerable concerns related to the legality of interferences with freedom of expression. Extremist content or material inciting violence is often difficult to identify, even for a trained human, due to the complexity of disentangling factors such as cultural context and humor. Algorithms are not currently capable of detecting irony or critical analysis. The filtering of speech to eliminate harmful content through algorithms therefore faces a high risk of over-blocking and removing speech that is not only harmless but might contribute positively to the public debate. On the other hand, the capacity of media content platforms to disseminate messages in real time and at a global scale substantially magnifies the reach, scope and thus the impact of harmful speech. The turn to automated approaches to on-line content filtering highlights the acute responsibility challenges which the increasing reliance on algorithmic systems in contemporary life generates: while they offer the benefits of scale, speed and efficiency relative to human decision-making, digital platforms claim that human oversight is necessarily inadequate, generating a 'responsibility gap' which they typically argue they cannot fairly be expected to fill.[81]

---

78 See Menn and Volz 2017

79 This principle is enshrined in EU-law and in relevant Council of Europe policy guidelines, including the recent Council of Europe CM/Rec(2018)2. See also UN General Assembly (2018). Several states have introduced laws or initiated law reform initiatives to address the spread of harmful on-line content. For example, Germany adopted its Network Enforcement Act ('NetzDG') in 2017. This law requires online platforms with more than two million registered users in Germany to remove 'manifestly unlawful' content, which contravenes specific elements of the German criminal code, such as holocaust denial and hate speech, within 24 hours of receiving a notification or complaint, and to remove all other 'unlawful' content within seven days of notification. Non-compliance risks a fine of up to €50 million. This law also seeks to increase platform responsibility through imposing greater transparency and significant reporting obligations. The law has been subject to significant criticism on the basis of its restrictive implications for freedom of expression. Eg Access Now 2018: 22. The UK has recently issued its *Online Harms White Paper*, which introduces a legal duty of care to make companies take more responsibility for the safety of their users and tackle harm caused by content or activity on their services, to be enforced by an independent regulator: UK Government 2019.

80 See Wagner Study 2017, 19.

81 See discussion of the so-called 'control' problem at Section 3.2.2 below.

(c) Right to privacy and data protection: Article 8

The Article 8 right to respect for private and family life and rights to data protection are being placed under unprecedented strain due to the ability of algorithms to facilitate the collection and repurposing of vast amounts of data, including personal data gleaned from digital observation of individual users which may generate further data, with entirely unpredictable results for the data subject.[82] As the Wagner Study observed, the use of personal data for the purposes of individual profiling, and its subsequent repurposing, threatens a person's right to 'informational self-determination'[83] particularly given that (as noted in section 2.1) even fairly mundane, innocuous data collected from the digital traces of individuals may be merged with other data sets and mined in ways that can generate insight that can enable quite intimate personal information to be inferred at a very high level of accuracy.[84] While contemporary data protection regimes (including Conv. 108 as modernised) are an important safeguard, conferring a set of 'data protection rights'[85] on data subjects, aimed at protecting them from unnecessary and unlawful data collection and processing, they might not provide comprehensive and practically effective guarantees against the use of intrusive profiling applications.

(d) The prohibition of discrimination in the enjoyment of rights and freedoms: Art 14

The potential for bias and discrimination arising from the use of machine learning (ML) techniques has attracted considerable attention, from both policy-makers and AI researchers alike. Concerns about unfair or unlawful treatment directly implicate Article 14 ECHR which provides that the enjoyment of the rights and freedoms set out in the Convention shall be 'secured without discrimination on any grounds such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status'.[86] There are many opportunities for bias to inadvertently affect the outputs produced by the use of machine learning techniques, arising from biases of the algorithms' developers, bias built into the model upon which the systems are generated, biases inherent in the data sets used to train the models, or biases introduced when such systems are implemented in real-world settings.[87] Not only might biased ML systems lead to discrimination and generate erroneous decisions, but this can entail significant wrongdoing, resulting in decisions that are systematically biased against groups that have historically been socially disadvantaged (and against individuals who are members of those groups), thereby reinforcing and compounding discrimination and structural disadvantage, even though these effects were not intended by the system's designers.[88] These concerns have been particularly

---

82 See, for example, tension between competition in on-line services and consumer privacy: Oxera 2018.

83 Wagner Study 2017:14.

84 Kosminski et al 2015.

85 The new rights introduced by the recently modernised Conv 108 include: the right not to be subjected to a decision significantly affecting him or her based solely on an automated processing of data without having his or her views taken into consideration, the right to obtain knowledge of the reasoning underlying data processing where the results of such processing are applied to him or her, and the right to object at any time, on grounds relating to his or her situation, and to the processing of personal data concerning him or her, unless the controller demonstrates legitimate grounds for processing which override his or her interest or rights in fundamental freedoms: Article 5 of the Modernised Convention 108.

86 Protocol No 12 ECHR Article 1 provides that 'the enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.' See also Art 21 CFEU.

87 Veale and Binns 2017.

88 Barocas and Selbst 2016; Wagner Study 2017: 27-28.

acute in relation to the use of machine learning techniques to inform custody and sentencing decisions within the US criminal justice system, due to allegations that such techniques operate in ways that are substantially biased against black and other racial minorities.[89] In response to these concerns, a growing body of work concerned with devising technical approaches for countering such bias has emerged.[90]

2.1.2 Societal risks associated with data-driven profiling

Contemporary applications of data-driven profiling technologies may also undermine important collective interests and values, only some of which fall within the scope of existing human rights protection. Much of the value of these technologies lies in their capacity to sort individuals and groups within a population, to automate decision-making, and to enable personalised, predictive interventions to be scaled and applied at the population-level**.** The following practices may generate significant societal risks yet these are often overlooked in public and academic debate.

a. **Population-wide, highly granular surveillance**

Because data-driven profiling requires the collection of highly granular data from individuals on a population-wide basis (i.e. at scale) to profile individuals and groups within and across a population to identify their inferred preferences and interests,[91] this necessitates the use of mass surveillance, often in a highly intrusive yet largely invisible manner. Although the threats which these practices pose to individual privacy and rights to data protection are readily apparent (discussed above), these practices also pose serious risks to the *collective* nature of privacy – thereby eroding the fundamental societal conditions in which individual privacy is possible and without which individual privacy cannot exist. As the Council of Europe's Parliamentary Assembly[92] observes,

> 'since many technologies nowadays can operate from a distance, most of us are not even aware of this mass surveillance and people are rather defenceless, since there are few possibilities to escape these surveillance activities. This creeping development and its impact on society and human rights have received so far little attention in political and public debate….(Yet) there has been little debate about the cumulative effects of mass surveillance. Instead, triggered by specific applications and incidents 'mini debates' have been organised, and the outcome of each debate is a balancing act that mostly favours national security or economic interests. The sum of the debates, however, is the gradual but steady dissolving of the privacy and anonymity of the individual'.[93]

---

89 Angwin et al 2016. But see Dieterich et al 2016.

90 See below at Section 3.7.1. As UN Special Rapporteur David Kaye has observed, 'Tackling the prevalence of discrimination in artificial intelligence systems is an existential challenge for companies and governments; failure to address and resolve the discriminatory elements and impacts will render the technology not only ineffective but dangerous.' UN General Assembly 2018, 18.

91 As the Council of Europe, Parliamentary Assembly's Committee on Culture, Science, Education and Media has observed at para 18 'The primary business model of the internet is built on mass surveillance': Council of Europe 2017.

92 Council of Europe 2017, para 60-61.

93 The Wagner Study also draws attention to the risks created by data aggregation and the generation of new data, which 'may then be mined through the use of algorithms, which creates a risk of large-scale surveillance ('data-veillance') by private entities and governments alike…a view echoed by the UN Human Rights Council (22 March 2017)': Wagner Study 2017: 15-16. As the Rathenau Instituut observes, 'modern-day surveillance via the IoT or internet, performed by states or companies, inherently involves the

These risks have magnified and deepened as a result of recent advances in AI capabilities that have fuelled the emergence of powerful biometric applications that can be used for identification purposes in ways that seriously threaten several human rights, including those protected under Article 8. In China, for example, AI-driven facial recognition technology is now being introduced in the Beijing subway to enable the facial features of subway users to be identified and tracked as they travel. These technologies have already been deployed in train stations, used at a pop concert to locate a suspected fugitive, and even implemented in schools to monitor student distraction and automatically alert the teacher when distraction is detected.[94] Nor is it difficult to imagine how powerful AI-driven lip-reading technologies recently developed by DeepMind (which are reported to outperform professional lip-readers[95]) could be deployed by repressive regimes in ways that magnify anxieties that strike at the very heart of the right to be left alone, and the potentially severe chilling effects that they may have on freedom of expression, individual self-development and democratic freedom, particularly when deployed by states to identify and detain individuals identified as political dissidents.[96] When combined with the use of data-driven profiling technologies that enable fairly innocuous and mundane data to be merged and mined in ways that may reveal highly personal characteristics (such as sexual orientation)[97] these can be very powerful tools in the hands of governmental regimes, whether liberal or repressive, and therefore generate acute threats to the exercise of all human rights and fundamental freedoms.

b. **Population-wide personalisation**

The attractions of profiling technologies are readily identifiable: for those wishing to engage in profiling, they enable the automated sorting and targeting of candidates of interest in order to *personalise* the way in which those individuals are treated. These techniques can be applied at scale, yet in ways that allow for real-time readjustment and reconfiguration of personalised offerings in response to user behaviour.[98] The capacity to engage in population-wide personalisation of digital services has potentially profound implications for social solidarity and community. Consider, for example, the practice of 'personalised pricing' that data-driven profiling and the rise of digital retailing makes possible. Under industrial capitalism, goods were mass produced and supplied to retailers, and typically made available to consumers at geographic locations in-store and on terms that applied universally to all customers entering the store at a particular time at the same price. In contrast, data-driven profiling now enables goods and services to be offered to potential customers at 'personalised' prices (because each customer only sees his or her own individualised 'digital shop front', and does not have access to the prices or offers made to others on-line), the level of which can be set by the use of data-driven profiling in order to identify the maximum 'willingness to pay' of each individual, thereby optimising revenue for the retailer.[99] While this kind of intentional discrimination might not be unlawful, in so far as it might not directly or indirectly discriminate individuals on

---

processing of personal data. Researchers are still trying to grasp the full extent of the harmful effects on the lives of individuals caused by such surveillance. The known effects are not comforting. Not only does surveillance have a chilling effect on speech…but it also leads to behavioural effects. For instance, as a result of surveillance, individuals conform to perceived group norms. This conforming effect occurs even when people are unaware that they are conforming (Kaminski & Witnov 2015). Both states and companies reinforce each other in their surveillance activities, as part of the *surveillance-innovation* complex (Cohen 2016)': Van Est and Gerritsen 2017: 20.

94 Cowley 2018.

95 Hutson 2018.

96 Donahoe 2016.

97 Kosinski et al, 2013.

98 Yeung 2016.

99 Townley et al 2017; Miller 2014.

the basis of protected grounds under contemporary equality law, nevertheless the effect is a serious departure from the pricing practices that prevailed in a pre-digital, pre-data driven age in ways that, if they become widespread and ubiquitous, may seriously undermine social solidarity and cohesion.[100]

c. Population-wide manipulation

The personalisation of informational environments that data-driven profiling makes possible brings with it new capacities to manipulate individuals in subtle but highly effective ways.[101] At the individual level, manipulation may threaten personal autonomy and an emerging right to cognitive sovereignty[102] but, as the recent Cambridge Analytica scandal in the run up to the US 2016 election and the Brexit referendum vividly illustrates, when deployed at scale for the purposes of political microtargeting to manipulate voting behaviour (which may entail the use of automated bots operating on social media websites), it may threaten the right to freedom of expression and information (Article 10) and could seriously undermine the foundations of democratic orders by perverting the right to free elections protected under Article 3, Protocol 1 ECHR.[103] The manipulative practices which so-called 'persuasive' digital technologies enable can be understood as interfering with rights protected under Articles 8 and 10 because they can be configured automatically (and continually reconfigured) to tailor the informational choice environment and architecture of individuals through the use of data-driven profiling to predict (often with great accuracy) the behaviours, interests, preferences, and vulnerabilities of individuals at scale. These applications can be used to manipulate and deceive individuals thus interfering with both informational and decisional privacy.[104]

The capacity to engage in manipulative practices has been exacerbated by the recent emergence of powerful AI applications that can simulate human traits (including voice simulation, visual representations of human behaviour and robots capable of interacting with humans with apparent emotional sensitivity), with such accuracy and precision that it can be extremely difficult for humans to detect that those traits are artificially generated. These technologies are likely to be attractive tools for malign actors to deceive and manipulate others. For example, some researchers already predict that advanced human-like synthesised voices will be used to gather information over the phone for deceptive and fraudulent purposes. If such attacks become commonplace and widespread, and cannot be readily detected by targeted individuals, this may seriously threaten the Article 5 right to liberty and security, and the collective security and respect for the rule of law upon which our individual and collective liberty and security depends. Opportunities to utilise these technologies to

---

100 Yeung 2018a. The European Commission has studied the prevalence of on-line personalisation practices: European Commission (2018b) In the UK, the Competition and Markets Authority (CMA) has commissioned economic research on the use of pricing algorithms and potential competition concerns, including collusion and personalised pricing: UK Competition and Markets Authority 2018.

101 Yeung 2016. For example, a recent study by the Norweigan Consumer Council analysed a sample of settings in Facebook, Google and Windows 10, and show how default settings and dark patterns, techniques and features of interface design meant to manipulate users, are used to nudge users towards privacy intrusive options: ForbrukerRadet 2018.

102 There is some academic support for recognition of a new right to 'cognitive sovereignty' aimed at providing individuals with rights-based protection against the forms of manipulation and deception that advancing digital technologies increasingly make possible, in order to guarantee to individuals a threshold level of sovereignty over their own minds (see Bublitz 2013). While this might be a self-standing right, it is also possible that such a right might be recognised as falling within Article 9(1) ECHR, which establishes the right to freedom of thought, conscience and religion.

103 Gorton 2016; Wagner Study 2017: 17. UK House of Commons, Digital Culture Media and Sport 2019.

104 Yeung 2016; Lanzing 2018; Council of Europe 2017.

undermine the integrity of the legal process might also become possible. As Brudage et al observe in their report on malicious AI:

> At present, recording and authentication technology still has an edge over forgery technology. A video of a crime being committed can serve as highly compelling evidence even when provided by an untrustworthy source. In the future, however, AI-enabled high-quality forgeries may challenge the 'seeing is believing' aspect of video and audio evidence. They might also make it easier for people to deny allegations against them, given the ease with which the purported evidence might have been produced. In addition to augmenting dissemination of misleading information, the writing and publication of fake news stories could be automated, as routine financial and sports reporting often are today. As production and dissemination of high-quality forgeries becomes increasingly low-cost, synthetic multi-media may constitute a large portion of the media and information ecosystem.[105]

d. Systematic treatment of individuals as objects rather than moral agents

Although the personalisation of individuals' informational environments is portrayed by social media companies as enabling the provision of more 'meaningful' content, there are two characteristics of the underlying socio-technological system upon which these practices rely that tend to treat individuals as objects rather than moral subjects. Firstly, individuals are singled out, not on the basis of any causal theory, but simply on the basis of correlations in data sets. As a result, these systems typically do not provide any reasoned account to individuals explaining why they have been singled out for treatment of a particular kind. Secondly, their underlying logic and their processing operations are highly complex and opaque, in ways that are practically, and sometimes technically, incomprehensible (discussed above). In other words, because many contemporary ML systems are designed to capture, commodify and optimise value extraction in the interests of the system owner, by tracking and analysing from the digital traces of the daily behaviour of individuals, they are not primarily concerned with identifying the *reasons* why individuals behave in particular ways. Rieder therefore refers to commercial applications of these 'big data' techniques as offering 'interested' readings of reality[106] in contrast to the disinterested pursuit of knowledge that characterises the pursuit of scientific inquiry for academic purposes.[107] The net effect of these applications is that humans are increasingly treated as objects rather than moral subjects, to be sorted, sifted, scored and evaluated by technological systems in ways that appear starkly at odds with the basic right of all individuals to be treated with dignity and respect, and which lies at the foundation of all human rights and fundamental freedoms.[108] As the EU European Group of Ethics (2018) explains,

---

105 Brundage et al 2018: 46.

106 Rieder 2016.

107 Merton 1942.

108 Law enforcement applications of AI for individual profiling within the criminal justice system are especially troubling. As AI Now has observed, Axon is now offering free body camera technologies to any US police department following their acquisition of two machine vision companies. It reports that 'Axon's new focus on predictive methods of policing – inspired by Wal-Mart's and Google's embrace of deep learning to increase sales – raises new civil liberties concerns. Instead of purchasing patterns, these systems will be looking for much more vague, context-dependent targets, like 'suspicious activity'. Behind the appearances of technical neutrality, these systems rely on deeply subjective assumptions about what constitutes suspicious behaviour or who counts as a suspicious person' per AI Now 2017: 25. Thus, individuals become 'objects of suspicion' on the basis of data analysis which have no demonstrable causal basis.

> 'AI driven optimisation of social processes based on social scoring systems with which some countries experiment, violate the basic idea of equality and freedom in the same way caste systems do, because they construct 'different kinds of people' where there are in reality only 'different properties' of people. How can the attack on democratic systems and the utilisation of scoring systems, as a basis for dominance by those who have access to these powerful technologies, be prevented?'… Human dignity as the foundation of human rights implies that meaningful human intervention and participation must be possible in matters that concern human beings and their environment. Therefore, in contrast to the automation of production, it is not appropriate to manage and decide about humans in the way we manage and decide about objects or data, even if this is technically conceivable. Such an 'autonomous' management of human beings would be unwelcome, and it would undermine the deeply entrenched European core values.'[109]

At the same time, commercial applications of AI for profiling purposes have been accompanied by the use of population-wide experimentation on individuals through the use of A/B testing, yet without being subject to the supervisory research ethics oversight provided by academic institutions pursuant to the Declaration of Helsinki. The latter sets out core requirements for the ethical conduct of human subject research.[110] The widespread and routine use of these practices again reflects a belief that human users are merely objects ripe for experimentation, so that fundamental norms and institutional oversight mechanisms that are designed to safeguard and protect the dignity and rights of individuals are not applicable. As Julie Cohen has put it '[W]e, the citizens have been reduced to raw material – sourced, bartered and mined in a curiously fabricated 'privatised commons' of data and surveillance.'[111]

e. Summary of the threats posed by data-driven profiling technologies

Taken together, the cumulative effects of the above practices resonate with the concerns about profiling expressed by Korff in his report for the Council of Europe concerning the trends, threats and implications for private life and data protection from the use of the internet and related services, which he expresses in the strongest possible terms. Because profiling systems provide the appearance of infallibility, objectivity, reliability and accuracy in the assessments that they produce, yet their outputs will inevitably and unavoidably generate errors (either false positives or false negatives) or generate discriminatory effects on certain groups[112] which are practically impossible for individuals to challenge, Korff concludes:

> 'Profiling thus really poses a serious threat of a Kafkaesque world in which powerful corporations and State agencies take decisions that significantly affect their customers and citizens, without those decision-makers being able or willing to explain the underlying reasoning for those decisions, and in which those subjects are denied any effective individual or collective remedies. That is how serious the issue of profiling is: it poses a fundamental threat to the most basic principles of the Rule of Law and the relationship between the powerful and the people in a democratic society.'[113]

These observations alert us to the collective and cumulative impacts of contemporary applications of data-driven technologies which, when undertaken systematically and at scale

---

109 European Group on Ethics in Science and New Technologies 2018: 9-10.

110 Kramer et al 2015; Tufecki 2015.

111 Powles 2015.

112 Korff and Browne 2013:6.

113 Korff and Browne 2013: 21.

may, over time, seriously erode and destabilise the social and moral foundations that are necessary for flourishing democratic societies in which individual rights and freedoms can be meaningfully exercised.

2.2 Collective societal threats and risks generated by other AI technologies

Although the concerns listed above can be attributed to the use of data-driven profiling, there are additional societal-level concerns and threats to collective interests and values that do not arise from the use of individual profiling. These include:

2.2.1 Malicious attacks, unethical system design or unintended system failure

Understandable and well-grounded fears have emerged concerning the safety and security implications of AI technologies, including concerns about the potentially catastrophic consequences of malicious attacks on AI systems (including data poisoning attacks and the use of adversarial ML) if safety critical systems are successfully targeted. But even if unintended, many fear the failure of AI technologies within safety-critical systems (such as autonomous vehicles) which could seriously harm public safety and security.[114] Worse, these systems could operate in ways that are designed to prioritise the safety of particular classes of persons over others, and which many would regard as unethical or even unlawful. As societies become increasingly dependent upon internet-enabled devices and cyber-physical systems more generally (many of which are safety critical), ensuring the safety and security of these systems acquires even greater importance.[115] This is especially due to the rise of various avenues and opportunities for malicious attack that are not confined to direct attack on the systems themselves, but may also include strategies aimed at exploiting network effects that enable the capacity to target and communicate to individuals at scale, yet with relative anonymity.[116]

2.2.2 Loss of authentic, real and meaningful human contact

In addition to above-mentioned concerns about the use of AI technologies to imitate human behaviour are diffuse but often deeply-felt anxieties that our collective life may become increasingly 'dehumanised', as tasks previously performed by humans are automated. Many fear that values and qualities that we cherish, including the value of real human interaction, of genuine empathy, compassion and concern, may be replaced by the relentless efficiency and consistency of AI driven services. These concerns are particularly prevalent when AI technologies are utilised in care environments (eg robot nurses, care nannies and other robotic care assistants) or in ways that otherwise threaten to denude our societies of characteristic values and features that inhere in real, authentic human contact, connection and relationships (such as the use of sex robots, for example) which, although inescapably fraught and imperfect, nevertheless contribute fundamentally to the meaning and value of human experience.[117] These applications have generated concerns about the need to ensure that they are designed and operate in ways that respect the dignity of those in care and might fall within

---

114 'All technologies are liable to failure, and autonomous systems will be no exception (which is pertinent to the issue of whether autonomous systems should ever be created without manual override)': Royal Academy of Engineering 2009: 3.

115 Thomas 2017a.

116 Brundage et al 2018. ForbrukerRadet 2018.

117 Yearsley 2017.

the scope of Article 8's protection of 'private and family' life and have prompted some to argue in favour of a 'right to meaningful human contact'.[118]

2.2.3 The chilling effect of data repurposing

Additional concerns arise from worries that people might refrain from participating in systems that could improve their life conditions (eg. seeking treatment for cancer) due to fears that personal data taken in highly sensitive contexts might be used by AI systems in other contexts in ways that may be contrary to their interests.[119] Concerns about these 'chilling effects' arising from the ease with which data obtained for one purpose may then be repurposed for other unrelated social ends helps explain the importance of honouring and upholding the 'purpose specification' principle enshrined in many contemporary data protection regimes. If individual autonomy and freedom is understood to include our capacity as individuals to move between multiple roles and identities, and to partition them and keep them separate if we so wish, then the systematic use of personal data for profiling and decision-making about individuals may threaten our capacity to do so.[120]

2.2.4 Digital power without responsibility

Worries that AI systems essentially treat people as objects rather than as moral subjects can be understood as part of a wider set of concerns about the exploitation of individuals in the service of so-called 'Big Tech'. There are several strands of concern. Firstly, there are serious concerns about the population-wide, instantaneous scale at which AI technologies can operate (eg Facebook News Feed) and the limited practical capacity for 'meaningful human oversight' of systems of this kind. The wedge between the capacity of machines relative to the capacity of humans to monitor them is evident in repeated claims by social media firms that they cannot realistically be expected to respect fully the rights of individuals by providing comprehensive, timely content moderation given the scale and speed at which their platforms operate, because – quite simply – they outpace human performance.[121] Yet by allowing AI driven automation to operate without comprehensive human oversight, this threatens to generate a serious responsibility gap, through which Big Tech reaps the benefits of these AI-driven platforms without the concomitant burdens.[122]

---

118 Concerns of this kind prompted the Parliamentary Assembly of the Council of Europe (PACE), to suggest that in contexts where human contact and interaction play a central role, as in raising children and caring for the elderly or people with disabilities, a 'right to meaningful human contact' could play a role: see Council of Europe Parliamentary Assembly 2017 para 65.

119 There is evidence that this 'chilling effect' has occurred in the US in that individuals have been unwilling to undertake genetic testing in circumstances where this would likely assist in their healthcare, owing to fears that the resulting information may be used be others in ways that will be contrary to their interest, particularly in employment and life insurance contexts: Farr 2016.

120 Understood in terms of Raz's conception of autonomy, which requires that individuals have an adequate range of options, then widespread data repurposing to inform organisational decision-making about individuals may effectively diminish our autonomy by reducing the range of options available to us. According to Raz, 'If a person is to be maker or author of his own life then he must have the mental abilities to form intentions of a sufficiently complex kind, and plan their execution. These include minimum rationality, the abiity to comprehend the means required to realize his goals the mental faculties necessary to plan actions etc. For a person to enjoy an autonomous ife he must actually use these faculties to choose what life to have. There must in other words be adequate options available for him to choose from. Finally, his choice must be free from coercion and manipulation by others, he must be independent.' Raz 1986: 373.

121 See above discussion at Section 2.1.1(b).

122 Keen 2018.

Not only does this constitute a violation of basic norms of social reciprocity, amounting to a kind of 'unjustified taking' from citizens and communities, but it entails the naked exercise of power without responsibility. In other words, the 'responsibility gap' which Matthias[123] claims has arisen from the emergence of computational systems with the capacity to learn[124] now has a more recent contemporary spin, at least in the context of social media platforms in which automated systems may be designed to remove or distribute content to users at a scale and speed which human content moderators cannot keep pace with, and which social media platforms claim that they cannot be responsible for.[125] Secondly, Big Tech has hitherto successfully managed to immunise themselves against external regulation by claiming to abide by 'ethical principles', which includes their claimed use of technological solutions (discussed in section 3.7.1) that seek to hard-wire normative values into the design and operation of technological systems but which, unless they are subject to external oversight and sanction, are unlikely to provide meaningful protection.[126]

2.2.5 The hidden privatisation of decisions about public values

AI technologies aim to reproduce or improve human performance with respect to some task that would require 'intelligence' if humans were to perform them. Yet the claim that these technologies 'outperform' humans is based on a very narrow definition of the overarching goal – couched in terms of performance of a narrowly defined task (such as identifying malign tissue from x-ray images). But in seeking to incorporate task-specific AI into complex socio-technical systems that are developed to provide services to individuals in real world contexts, this invariably implicates a wider range of values beside that of precision and efficiency in task performance.

These systems will invariably reflect the values and value priorities of the system and its developers and might not be aligned with the collective values of the public or the democratic and constitutional values that human rights are designed to serve. Yet, even in relation to AI systems that directly affect and interface with the public, citizens and other affected groups and organisations will typically not be given any meaningful opportunity to participate in identifying these values or value trade-offs that these systems are configured to reflect.[127] The use of ML in risk-scoring systems used to evaluate the 'recidivism risk' of convicted criminals seeking release from custody offers a vivid example: although the criminal justice system in contemporary democracies is founded on, and is expected to give effect to, several important criminal justice values, these scoring systems have hitherto been designed to optimise only with one such value: public protection.[128] As AI technologies become embedded as tools for optimising the efficiency of social coordination (such as smart navigation systems or smart infrastructure management, for example), they will inevitably make decisions that prioritise some values over others and impact directly on individuals and groups, some of whom may benefit and others who may not. Yet, as Sheila Jasanoff[129] and other STS scholars have repeatedly highlighted, technological systems reflect normative values. Given their widespread effects, the determination of those values should be subject to democratic

123 Matthias 2004.

124 Discussed below at section 3.3.2.

125 But see discussion at n.79 above.

126 See section 3.3.4 below. These technological strategies can be interpreted as "handing-off" human rights concerns to tech firms, empowering them to define (often with a narrow scope) the scope and content of a user right and over which they have exclusive powers of enforcement.

127 Korff and Browne 2013.

128 Zweig et al 2018.

129 Jasanoff 2016.

participation and deliberation rather than being resolved privately by private providers motivated by commercial self-interest.

2.2.6 Exploitation of human labour to train algorithms

AI and ML systems are often claimed to 'outdo human performance' because the algorithms are trained by large numbers of human workers. For example, an ML algorithm for answering search queries will be evaluated against an army of Mechanical Turk workers who act like the algorithm until the algorithm outperforms their answers. Even after the algorithm has been trained, there may be unwanted side effects of the use of these automated algorithms, requiring humans to identify and weed them out. This is exemplified in the case of social media content moderators who are asked to remove inappropriate content on social networks. Both the training for ML models, as well as the consequent human clean-up activities to weed out the models' externalities, are often concealed to maintain the mythology of seamless automation.[130] The humans who train ML models are often located in poor communities, often in the global south, and typically work under extremely precarious conditions.[131] Nor are they typically provided with support for dealing with the psychological burdens that may come with the 'clean up' activities. Some claim that because many ML algorithms continue to learn on their general user population, this allows the system owners to 'free ride' on user labour, thereby nurturing a mode of AI production that contributes to the creation of conditions in which unpaid labour is normalised and legitimised, while human workers are denuded of rights or recognition.[132]

2.3 Power asymmetry and threats to the socio-technical foundations of moral and democratic community

The above-mentioned adverse impacts arising from the increasing power and sophistication of new and emerging digital technologies are exacerbated by the radical asymmetry in power between those who develop and deploy algorithmic systems and the individual users who are subject to them. This asymmetry in power arises largely due to the former's unique ability to engage in synoptic, pervasive real-time surveillance of users, collecting and accessing massive data sets gleaned from users' digital interactions on a continuous and real-time basis. This, in turn, enables them to subject individuals and populations to algorithmic evaluation in order to 'sort and score' them accordingly[133] while empowering platform owners to communicate directly to users on a one-to-many basis automatically and at scale. In contrast, the practical capacity of individuals to understand and navigate the complexity of the data ecosystems in which they are embedded is extremely limited, as is individual users' ability to identify whether or not digital information and other services are being made available to them on the same terms as to other users.[134]

This power asymmetry suggests that, at least under current institutional arrangements, we need to re-examine the capacity of existing rights and our *existing* mechanisms of oversight and enforcement to respond comprehensively to the risks associated with our increasingly powerful digital technologies. As the Wagner Study has observed:

---

130 Irani 2015.

131 See for example Chen 2014.

132 Ekbia and Nardi 2014.

133 Ferraris et al 2013.

134 See the findings reported by *Which?* 2018. Mireille Hildebrandt refers to the 'digital unconscious' which is flooded with data, in contrast to the information individuals can connect with: Hildebrandt 2015: 196.

> the increasing use of automation and algorithmic decision-making in all spheres of public and private life is threatening to disrupt the very concept of human rights as protective shields against state interference. The traditional asymmetry of power and information between state structures and human beings is shifting towards an asymmetry of power and information between operators of algorithms (who may be public or private) and those who are acted upon and governed.[135]

In particular, existing human rights institutions may struggle to provide effective and meaningful protection for at least three reasons.

Firstly, given the highly complex and opaque nature of these technologies, it is very difficult in practice for individuals to identify if their rights have been infringed, and if so in what ways. Often individuals will be unaware that these technologies are being used for the purposes of evaluating them. Even if individuals are willing to assert their human rights against infringements that arise from the use of automated decision-making, for example, the remedies available to them might not provide them with their desired outcome. Perhaps, for example, it is not so much that individuals want an explanation of why they were treated less favourably than others, but they want to insist that they should be entitled to equally favourable treatment.[136]

Secondly, even if individuals are aware that their rights may have been interfered with as a result of the use of AI systems, one might question the likelihood that individuals will in practice seek to initiate remedial action if circumstances where they do not regard the interference as sufficiently serious to motivate them to invest the time, energy and resources associated with launching and maintaining a compliant. The resulting collective action problem means that the aggregate adverse impacts of these systems are likely to continue unremedied, at least in the absence of collective complaints mechanisms or an official body with the competence, resources and remit to take enforcement necessary to ensure effective human rights protection.

Thirdly, many of the larger adverse societal concerns cannot be readily expressed in the language and discourse of human rights because they concern *collective* values and interests, including threats to the broader and more amorphous moral, social and political culture and context in which advanced digital technologies operate. At the same time, the speed and scale at which these technologies now operate poses novel threats, risks and challenges which contemporary societies have not hitherto had to contend with. Yet many of the anticipated cumulative and collective effect of these systems over time could fatally undermine the social and technical conditions that are essential for the exercise of human rights and fundamental freedoms. Because current approaches to the interpretation and enforcement of human rights are highly individualised in orientation,[137] they are likely to struggle to address the *collective, aggregate and cumulative* risks and harms that these technologies might generate. In other words, existing rights-based approaches and rights discourse tend to overlook deeper systematic, societal concerns, including threats to the underlying democratic and moral fabric in which individual rights are anchored and without which they have no meaning.[138]

---

135 Wagner Study 2017: 33.

136 Edwards and Veale 2017.

137 Yeung 2011.

138 See section 3.8 below.

2.4 Summary

This section has examined the adverse individual and collective threats and risks to society that the application of advanced digital technologies may pose. It has emphasized the way in which the widespread and growing use of advanced digital technologies (including AI), particularly those which rely upon data-driven profiling technologies, may systematically threaten the exercise of human rights, as well as more general collective values and interests that fall outside the scope of existing understandings of human rights protection. It has also considered the threats and risks posed by other AI technologies and their contemporary and anticipated applications. These include concerns associated with hostile and malicious applications or the unethical or unsafe design and operation of AI-enabled systems, diminishing opportunities for authentic, real and meaningful human contact, the chilling effect of data repurposing, the exercise by digital platforms and others with AI capabilities of power without responsibility, the creeping yet hidden privatisation of decisions about public values, and the exploitation of human workers to train algorithms. Finally, it has highlighted the growing power asymmetry between those with the capacity and resources to develop and employ AI technologies and the individual users, groups and populations directly affected by their use, which may substantially diminish their capacity to identify and seek protection and redress under existing rights-protecting institutions. The wide-ranging and potentially serious individual and collective threats and risks associated with the development and application of advanced digital technologies inevitably raise important questions about how responsibility for avoiding, preventing, and mitigating them should be allocated. Furthermore, if those risks ripen into harm and/or violate human rights, how should responsibility for those consequences be attributed and allocated and what institutional mechanisms can be relied upon to ensure adequate enforcement and redress, particularly given the collective action problem faced by individual rights-holders? It is these questions that Chapter 3 seeks to address, beginning with an examination of the concept of responsibility, why responsibility matters and an analysis of the ways in which AI technologies challenge existing conceptions of responsibility.

Chapter 3. Who bears responsibility for the threats, risks, harms and wrongs posed by advanced digital technologies?

As the preceding section has demonstrated, advanced digital technologies generate serious threats and risks to our individual and collective interests and values and may perpetuate the commission of substantial and systematic wrongdoing, including human rights violations. Taken together, these threaten the health of the collective moral and social foundations of democratic societies. Accordingly, this section considers who bears responsibility for their prevention, management and mitigation, and for making reparation if they ripen into harms and rights violations, to individuals, groups and to society. The following discussion highlights how the concept of responsibility is implicated by the emergence of advanced digital technologies (including AI), particularly in light of their implications for human rights protected under the ECHR referred to in Chapter 2.

The following discussion proceeds in several stages.

Firstly, it begins by clarifying what we mean by responsibility and why responsibility matters, emphasising its vital role in securing and giving expression to the rule of law and which is essential for peaceful social co-operation.

Secondly, it then considers two core themes raised in contemporary discussions of the adverse risks associated with AI technologies, notably the role of the tech industry in promulgating and voluntarily committing themselves to abide by so-called 'ethical standards' and secondly, the alleged 'control problem' that is claimed to flow from the capacity of AI-driven systems to operate more or less autonomously from their human creators.

Thirdly, it identifies a range of different 'responsibility models' that could be adopted to govern the allocation of responsibility for different kinds of adverse impacts arising from the operation of AI systems, including models based on intention/culpability, risk creation/negligence, strict responsibility and mandatory insurance schemes. Because the focus of this report is on the implications for human rights, responsibility for human rights violations is widely understood 'strictly' (or as 'strict' – so that provided a human rights violation has been established, there is no need for proof of fault). In contrast, the allocation of obligations of repair for tangible harm to health or property may be legally distributed in accordance with a variety of historic responsibility models. Because the allocation of historic responsibility for tangible harm arising from the operation of AI systems also has a prospective dimension, through its guiding function in identifying the nature and scope of the obligations of those involved in the development, production and implementation of AI systems, these responsibility models are briefly outlined.

Fourthly, it draws attention to the acute challenges for the allocation of responsibility generated by the operation of complex and interacting socio-technical systems, which entails contributions from multiple individuals, organisations, machine components, software algorithms and human users, often in complex and highly dynamic environments.

Fifthly, it draws attention to a range of non-judicial mechanisms for securing both prospective and historic responsibility for the adverse impacts of AI systems, including various kinds of impact assessments, auditing techniques and technical protection mechanisms.

Sixthly, it emphasises the role and obligations of states in relation to the risks associated with advanced digital technologies, focusing specifically on their obligations to ensure effective protection of human rights.

Finally, it highlights the need to reinvigorate human rights discourse in a digital age, drawing attention to the need to protect and nurture the socio-technical foundations necessary for human agency and responsibility, without which human rights and freedoms cannot be practically or meaningfully exercised.

3.1 What is responsibility and why does it matter?

In setting out the aims of this study, we have already noted that a society's conceptions and practices of responsibility are of vital importance because they serve to ensure that, within constitutional democratic orders, individuals and organisations are held to account for the adverse 'other-regarding' effects of their actions. Despite the extensive legal and philosophical literature concerned with responsibility, relatively few academics focus their attention on the fundamental role of responsibility for individuals and for society. Woven beneath the surface of this scholarship lies recognition that the concept of responsibility serves two critical functions, broadly reflecting what moral philosopher Gary Watson's refers to as the 'two faces' of responsibility.[139] The first face is essential to our sense of 'being in the world' as moral agents, that is, as authors of our own lives who act on the basis of reasons. As Watson puts it:

> Responsibility is important to issues about what it is to lead a life, indeed about what it is to have a life in the biographical sense, and about the quality and character of that life. These issues reflect one face of responsibility (what I will call its aretaic face).[140]

But Watson identifies a second face of responsibility which is concerned with practices of holding people accountable.[141] For him,

> when we speak of conduct as deserving of "censure," or "remonstration," as "outrageous," "unconscionable" (and on some views, even as "wrong"), is to suggest that some *further response* to the agent is (in principle) appropriate. It is to invoke the practice of holding people morally accountable, in which (typically) the judge (or if not the judge, other members of the moral community) is entitled (in principle) to react in various ways.

The difference between these two faces of responsibility, which we might call the 'self-disclosure' view of responsibility on the one hand, and the 'moral accountability' view on the other, is illuminated in the following scenario:

> If someone betrays her ideals by choosing a dull but secure occupation in favor of a riskier but potentially more enriching one, or endangers something of deep importance to her life for trivial ends (by sleeping too little and drinking too much before important performances, for example), then she has acted badly—cowardly, self-indulgently, at least unwisely. But by these assessments we are not thereby holding her responsible, as distinct from holding her to be responsible. To do that, we would have to think that she is accountable to us or to others, whereas in many cases we suppose that such behavior is "nobody's business." Unless we think she is responsible to us or to others to live the best life she can—and that is a moral question—we do not think she is accountable here. If her timid or foolish behavior

139 Watson 2004.

140 Watson 2004: 262-263.

141 Watson 2004: 264.

> also harms others, and thereby violates requirements of interpersonal relations, that is a different matter.[142]

A similar sentiment is reflected in the concept of 'basic responsibility' articulated and developed by legal scholar John Gardner who claims that our basic responsibility is central to our sense of being in the world. It is fundamental to our identity as rational agents, that is, as creatures who act on the basis of reasons and who, as individuals, want our lives to make rational sense, to add up to a story not only of whats but also of *whys.*[143]

For Watson, control is arguably central to the accountability practices that characterize the second face of responsibility.

> Because some of these practices—and notably the practice of moral accountability—involve the imposition of demands on people, I shall argue, they raise issues of fairness that do not arise for aretaic appraisal. It is these concerns about fairness that underlie the requirement of control (or avoidability) as a condition of moral accountability. 'Holding responsible' can be taken as equivalent to 'holding accountable'. But the notion of 'holding' here is not to be confused with the attitude of *believing* (as in, 'I hold that she is responsible for *x*'). Holding people responsible involves a readiness to respond to them in certain ways. To be "on the hook" in these and other cases is to be liable to certain reactions as a result of failing to do what one is required. To require or demand certain behavior of an agent is to lay it down that unless the agent so behaves she will be liable to certain adverse or unwelcome treatment. For convenience, I shall call the diverse forms of adverse treatment "sanctions." Holding accountable thus involves the idea of liability to sanctions. To be entitled to make demands, then, is to be entitled to impose conditions of liability.[144]

Because this study is concerned with identifying where responsibility should lie for the individual and collective threats, risks, harms and human rights violations stemming from advanced digital technologies, it focuses primarily on the second face of responsibility, understood in terms of 'holding accountable'. Nevertheless, there is a crucial link between these two faces of responsibility that rests in the status of the individual as a *moral agent* with the capacity to make active choices and decisions, including decisions that affect, and have the potential to cause harm or to perpetuate wrongs to others. As Gardner puts it, '(w)e are moral agents only insofar as we are basically responsible.'[145] Basic responsibility is therefore central to, and a reflection of, both faces of responsibility. As Gardner observes, whenever we perpetrate wrongs or mistakes, we always hunt around for justifications and excuses, not only because, as rational beings we want to avoid (unpleasant) consequential responsibility (the 'Hobbesian explanation') but also for a deeper reason (which he refers to as the 'Aristotelian explanation') that, as rational beings, we all want to assert our basic responsibility – and this requires that I can give a good account of myself.[146]

Responsibility and the rule of law

In other words, basic responsibility is essential, not only for our self-understanding as individuals as authors of our own lives but also as individuals *as members of a community of*

---

142 Watson 2004: 265-266.

143 Gardner 2003.

144 Watson 2004: 272-273.

145 Gardner 2008: 140.

146 Gardner claims that this account that we provide need not be to anyone in particular, but can and should be offered to all the world. Hence he rejects the view that responsibility is necessarily 'relational'.

moral agents. Moral agents have the capacity and freedom to make choices about their decisions and actions, and to do so in ways that might be wrongful or cause harm, whether to other individuals or to the conditions that are essential to maintain the stability and social co-operation needed to sustain community life. It is our basic responsibility and our responsibility practices through which members of a community hold each other to account, that characterise a political community largely as a *moral* community (i.e. a community of moral agents). Of critical importance is the mutual respect and self-restraint exercised by members of a moral community that makes possible and sustains community life, and which ultimately lies at the foundations of the contemporary rule of law ideal.[147] A society that lacks a system for institutionalising its responsibility practices in order to hold people responsible for the adverse impacts of their other-regarding conduct (including conduct that harms others or violates their human rights) would not benefit from the vital protective functions that such an institutional system provides and that are essential for peaceful and trustworthy social co-operation and coordination. In other words, our system for ensuring that responsibility is duly allocated plays a critical role in sustaining the underlying social framework of cooperation without which the law cannot rule. At the same time, it is important to recognise that the stability and continuity of these social foundations rest, ultimately, on the mutual respect and self-restraint of individual members of the moral community and not on a system of technological coercion and control. It is this mutual respect and self-restraint that is absent from the ostensibly happy, stable, orderly and efficient society depicted in Huxley's *Brave New World.*[148] Inhabitants of *Brave New World* have no meaningful rights or freedoms. Theirs is not a moral community but a society comprised of members who are merely passive objects, whose thoughts and actions have been hard-wired and controlled by and through the exercise of technological power of an authoritarian dictator, and in which notions of freedom, autonomy and human rights not only fail to flourish, but simply lack any meaning or purchase.[149]

Accountability, answerability and transparency

The critical importance of institutionalised systems of responsibility to secure the social foundations upon which the rule of law is founded highlights the need, within any moral and political community committed to respect for human rights, to establish and implement institutional mechanisms for holding members of the community to account for their other-regarding conduct. Although the concept of accountability is contested, for present purposes it has been usefully described as 'requiring a person to explain and justify – against criteria of some kind – their decisions or acts, and then to make amends for any fault or error.'[150] So understood, accountability mechanisms possess the following four features: setting standards against which to judge the account, obtaining the account, judging the account, and deciding what consequences (if any) should follow. The concept of accountability is of particular importance in relationships between a principal and agent, in which the agent is expected to act for and on behalf of the principal, who is therefore required to give an account – to be *answerable to* - the principal on whose behalf the agent acts. Transparency is directly linked to accountability, in so far as accountability requires that those being called upon to account can explain the reasons for their actions, and to justify those actions in accordance with a particular set of rules or standards for evaluation. Transparency is therefore important for at least two reasons: to enable those affected by a decision or action to know the reasons for

147 Galligan 2006.

148 Huxley 1932; Yeung 2017b.

149 Yeung 2011.

150 Oliver 1994: 245. See also Bovens 2007 and literature cited therein.

that action or decision, and to enable the affected party to evaluate the quality of those reasons.[151]

Mechanisms for accountability have particular importance in relation to the exercise of governmental power within liberal democratic societies because governmental officials are regarded as the servants of the citizens upon whose behalf they act and from whom their power is ultimately derived. Yet, the importance of accountability arises whenever the exercise of power has the capacity to affect others in adverse ways. Accordingly, concerns about the power, scale and effects of complex socio-technical systems that rely upon AI technologies have given rise to a cluster of concerns that can be understood as united in a concern to secure 'algorithmic accountability', particularly given the opacity of these systems and their potential to be utilised in ways that can have highly consequential implications for individuals, groups and society in general.[152] Securing accountability and responsibility for human rights violations and other adverse consequences resulting from the operation of these technologies is therefore essential. Although existing laws, including data protection law, consumer protection law, competition law and constitutional laws that enshrine legal protection for human rights within national legal systems, have the potential to play a significant and important role in securing various dimensions of algorithmic accountability, their contribution to securing algorithmic accountability is beyond the scope of this study. Rather, the following discussion seeks to examine implications of advanced digital technologies (including AI systems) for the *concept of responsibility*, focusing primarily on their implications for human rights violations, drawing on both moral philosophy and legal scholarship.

3.2 Dimensions of responsibility

This general concept of responsibility as 'holding accountable' has been extensively examined in the legal and philosophical literature, and various insights from that literature are selectively drawn upon in the analysis which follows. Although there are many different 'senses' in which the term 'responsibility' is used,[153] for the purposes of this study, the temporal element of responsibility is worth emphasising, facing in two directions:

(a) Historic (or retrospective) responsibility: which looks backwards, seeking to allocate responsibility for conduct and events that occurred in the past. As we shall see, considerable difficulties are claimed to arise in allocating historic responsibility for harms and wrongs caused by AI systems; and

(b) Prospective responsibilities: which establish obligations and duties associated with roles and tasks that look to the future, directed towards the production of good outcomes and the prevention of bad outcomes. Prospective responsibilities serve an important guiding function. As Cane puts it, 'one of the most important reasons why we are interested in responsibility and related concepts is because of the role they play in practical reasoning about our rights and obligations vis-à-vis other people, and about the way we should behave in our dealings with them.'[154] In the context of responsibility for the actions and resulting consequences of

151 Yeung and Weller 2018b. Zalnieriute et al 2019.

152 Yeung 2017.

153 Hart 1968: 211-230.

154 Cane 2002: 45.

autonomous AI/robotic systems, the idea of 'role responsibility'[155] has sometimes been foregrounded.[156]

Any legitimate and effective response to the threats, risks, harms and rights violations posed by advanced digital technologies is likely to require a focus on the consequences for individuals and society which attends to, and can ensure that, *both* prospective responsibility aimed at preventing and mitigating risks, and historic responsibility for adverse effects arising from the operation of the complex socio-technical systems in which these technologies are embedded, is duly and justly assigned. Only if both the historic and prospective dimensions of responsibility are attended to can individuals and society have confidence that efforts will be made first, to prevent harms and wrongs from occurring, and secondly, if they do occur, then institutional mechanisms can be relied upon to ensure appropriate reparation, repair and to prevent further harm or wrongdoing. It will necessitate a focus on both those involved in the development, deployment and implementation of these technologies, individual users and the groups affected by them and action by the state (and states acting collectively and cooperatively) to ensure the establishment and maintenance of conditions needed to safeguard citizens against unacceptable threats and risks, thereby ensuring that human rights are adequately protected. In other words, proper consideration of the responsibility of AI technologies and systems will attend to the positions of both the moral agent and the moral patient, as well as the larger moral community more generally, in order to answer the questions: *responsibility to whom and for what?* [157]

3.3 How do advanced digital technologies (including AI) implicate existing conceptions of responsibility?

Having clarified what we mean by responsibility and highlighted the need to attend to both its prospective and retrospective dimensions, we are now in a position to consider where responsibility lies for the adverse consequences, threats and risks associated with the development and implementation of AI technologies, including human rights violations and other wrongs and harms arising from their operation. Although this question is simple to state, there are considerable conceptual challenges in seeking to answer it. As the EU European Group on Ethics[158] has observed, AI technologies raise:

> …questions about human moral responsibility. Where is the morally relevant agency located in dynamic and complex socio-technical systems with advanced AI and robotic components? How should moral responsibility be attributed and apportioned and who is responsible (and in what sense)?

In other words, the complexity of the technologies themselves, and the larger socio-technical contexts in which they are implemented and applied, can obscure lines of moral responsibility, particularly when they operate in unexpected ways that generate harm or violate rights. But we must bear in mind that moral responsibility and legal responsibility are distinct, albeit related, concepts. Unlike morality, the law has a highly developed system for institutionalizing and enforcing responsibility (including the application of sanctions in certain circumstances) because it must adjudicate real world disputes, and which requires both finality of judgement and legal certainty.[159] A society cannot rely exclusively on individuals' inclinations to 'act

---

155 Hart 1968: 211-230.

156 See Section 3.3.1 below.

157 Liu and Zawieska 2017; Cane 2002.

158 EU European Group on Ethics 2017.

159 Cane 2002.

ethically' because the lack of any institutional mechanisms to enforce those standards (including lawful authority to sanction non-compliance) means that such an entirely voluntary system for would fail to provide the stable and reliable social foundations necessary for trustworthy and peaceful social cooperation within contemporary societies. The law's role in securing and institutionalising responsibility to ensure the protection of legal rights and enforce the performance of legal duties is therefore essential. As the following discussion demonstrates, the way in which legal systems have allocated historic responsibility has typically been more sensitive to the interests of victims and of society in security of the person and property in comparison with moral philosophical accounts of responsibility, which have tended to focus on the conduct of the moral agent and whether it appropriately attracts blame. Yet applying these moral and legal concepts of responsibility to the development and implementation of advanced digital technologies (including AI) in contemporary contexts may not be straightforward. The capacity of these technologies and systems to operate in ways that were not previously possible may challenge our existing legal, moral and social conceptions of responsibility, particularly given the properties identified in section 2.1 above as responsibility-relevant, including their:

- inscrutability and opacity
- complex and dynamic nature
- reliance on human input, interaction and discretion
- general purpose nature
- global interconnectivity, scalability and ubiquity
- automated, continuous operation, often in real-time
- reliance on large data-sets
- capacity to generate 'hidden' insight from merging data sets
- ability accurately to imitate human traits
- greater software complexity (include vulnerability to failure and malicious attack), and
- capacity to 'personalise' and configure individual choice environments
- capacity to redistribute risks, benefits and burdens among and between individuals and groups via the use of AI-driven optimisation systems which reconfigure social environments and choice architectures, and
- capacity to generate collective action problems.

Before proceeding, it is important to clarify the conceptual distinction between two different types of adverse effects that may (and have) arisen from the operation of AI systems:

(a) violations of **human rights,** including but not limited to, the rights protected under the ECHR;
(b) **tangible harm** to human health, property or the environment;

These are separate and distinct concepts and consequences. It is possible for a human rights violation to occur without any tangible harm, and vice versa. For example, the removal by Facebook in 2016 of the iconic photograph of a naked 9-year old girl fleeing napalm bombs during the Vietnam War, on the grounds that nudity violated its community standards, can be understood as a violation of the Article 10 right to freedom of expression and information, although it did not generate any substantial tangible harm.[160] Conversely, if a self-driving car collides with and injures a wild animal, this entails the infliction of harm without any human rights violation. Yet any given event or series of events may entail both tangible harm and a violation of human rights. Thus, if a self-driving vehicle collides with and fatally injures a

[160] See Scott and Isaac. 2016

pedestrian, this would entail both a violation of the Article 2 right to life and the infliction of tangible harm.[161]

The focus of this report is on examining the responsibility implications of AI systems from a human rights perspective. It is therefore primarily concerned with analysing responsibility for human rights violations, rather than responsibility for tangible harm arising from the operation of these systems. The following discussion focuses primarily upon those who create, develop, implement and preside over AI systems. It asks whether they can be held responsible for the adverse consequences those systems might generate, beginning with an examination of two core themes that have arisen in contemporary responses concerned with identifying where responsibility lies for the risks which AI technologies may pose: first, voluntary action by the tech industry in promulgating and publicly proclaiming their commitment to so-called 'ethical guidelines', and secondly, claims that because AI systems act autonomously, this relieves their creators from responsibility for their decisions and any consequential adverse effects. The obligations of the state in relation to these adverse effects is considered after various 'models of responsibility' that might apply in ascribing responsibility to those who develop and implement AI systems have been described.

3.3.1 Prospective responsibility: voluntary ethics codes and the 'Responsible Robotics/AI' project

Rising public anxiety and the recent 'Techlash'[162] in response to the growing power, practices and policies of the Big Tech firms, particularly following the use of political micro-targeting and the Cambridge Analytica scandal, have precipitated numerous voluntary 'ethics' initiatives by the tech industry. These initiatives typically entail the promulgation of a set of norms and standards either by individual tech firms or by a group of tech firms (including non-profit organisations[163] or a technical standard setting organisation[164]) publicly and voluntarily espousing their commitment to comply with those publicised standards of conduct (often called 'codes of ethical conduct').[165] These initiatives can be understood as part of a movement towards what Liu and Zawieska refer to as the 'responsible AI/robotics' project.[166]

Two features of these initiatives are worth highlighting. Firstly, they are concerned with *prospective responsibility*, seeking to identify and allocate 'role responsibility' (or spheres of obligation) for those involved at each stage of the design, development and deployment of these technologies with the aim of demonstrating to the public the seriousness of their commitment to addressing ethical concerns.[167] One notable feature of these initiatives is that they tend steadfastly to avoid explicitly referring to the *historic responsibilities* of those involved in the design, development and deployment of these technologies when things go awry. Neither do they tend to specify upon whom the *blame* should fall for such consequences, nor acknowledge any *obligation to compensate* those adversely affected.

---

161 The scope of adverse effects regarded as constituting legally recognisable 'harm' varies between national legal systems. In Anglo-Commonwealth common law systems, for example, some forms of non-tangible harm (such as emotional distress and mental anguish) may be legally recognised as harm for the purposes of compensation awards in personl injury cases: Gilliker 2000.

162 The Economist 2018b.

163 For example, the 'beneficial AI' movement is supported by the Future of Life Institute: see Conn 2017.

164 See for example, the various recommendations and guidelines developed by the IEEE's Global Initiative for Ethical Considerations in AI and Autonomous Systems 2017.

165 For example, Google's 'Objectives for AI Applications', see Pichai 2018.

166 Liu and Zawieska 2017.

167 Liu and Zaweiska 2017; Loui and Miller 2007; Eschelman 2016.

Rather, as Liu explains, role responsibility describes 'a sense of responsibility that attaches to an individual by virtue of the position he/she occupies or the function that he/she is expected to fulfil and is therefore by the performance of obligations connected to an individual's role and which can be pre-defined and specified in advance.'[168] Thus, once an individual has discharged the duties attached to his or her role or office, that is regarded as due fulfilment of his or her responsibilities.[169]

Secondly, these 'Responsible AI/robotics' initiatives can be characterised as an emerging professional self-governance movement which can be located within a longer standing social phenomena often discussed under the rubric of 'corporate social responsibility'. The character of these so-called 'ethical codes' as 'social' (rather than legal) and entirely voluntary, means that they the obligations and commitments specified in these codes are not legally enforceable if violated. Nor do these initiatives typically make provision for the establishment and maintenance of enforcement institutions and mechanisms through which an independent, external body is empowered to evaluate the extent to which those commitments have been complied with or to impose sanctions for non-compliance. Thus, although these initiatives provide welcome recognition by the tech industry that the ethical development and deployment of advanced digital technologies is a matter of public concern that warrants their action and attention, these initiatives lack any formal institutional mechanisms to enforce and sanction violations. Nor is there any systematic representation of the public in the setting of those standards. Accordingly, these initiatives have been roundly criticised as a form of 'ethics washing'[170] failing to take ethical concerns seriously.[171]

If these codes of practice were supported by institutional mechanisms, *backed by law*, including provision for external participation in the setting and evaluation of the standards themselves, and independent, external oversight to evaluate whether individual firms and organisations have in fact complied with the specified norms and standards, there would be stronger basis upon which those affected (and society more generally) could have confidence that *meaningful* and *democratically legitimate* safeguards were in place to prevent and mitigate some of the ethical risks associated with these technologies (see section 3.7 below).[172] It is the need for meaningful and effective safeguards that a human rights perspective insists upon.[173] At the same time, prospective approaches cannot ensure that *historic responsibility* in the event that harm or wrongdoing occurs will be duly allocated. As Liu and Zaweiska argue, although the 'Responsible Robotics/AI' project may be welcomed, it leaves a 'responsibility gap', because it is only concerned with role responsibility rather than causal responsibility. Unlike role responsibility, causal responsibility is a form of historic responsibility. Its concern is to identify and establish a relation between cause and effect. It is thus *retrospective* in nature, inherently outward-looking, relational in orientation, because it foregrounds the moral patient

---

168 Cane is critical of the narrowly defined way in which role responsibility is attached to specific roles or tasks, observing that 'being a responsible person involves taking seriously the prospective responsibilities, whatever they are, attaching to whatever activity one is engaged in at any particular time': Cane 2002: 32.

169 Liu 2016: 336.

170 Wagner 2019. Metzinger 2019.

171 Green et al 2019; Hagendorff 2019.

172 Nemitz 2018.

173 See AHRC supra, n.10. As David Kaye, Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression to the UN General Assembly stated, 'The development of codes of ethics and accompanying institutiona structures may be an important compement to, but not a substitute for, commitments to human rights. Codes and guidelines issued by both public and private sector bodies shoud emphasize that human rights law provides the fundamental rules for the protection of individuals in the context of artificial intelligence', UN General Assembly 2018: 18.

(that is, the person or persons harmed by the relevant activity)[174]. In contrast, the allocation of role responsibility focuses on the prospective role responsibilities of those identified as responsible agents. Accordingly, a 'responsibility gap' arises, because discharging one's prospective or role responsibilities will not necessarily guarantee that causal responsibility will be duly allocated.[175] In other words, the designation of role responsibility cannot ensure retrospective accountability nor allocate blame because it is concerned only with the fulfilment of pre-established obligations, rather than atonement and accountability for consequences.[176]

3.3.2 Machine autonomy and the alleged 'control' problem

(a) The alleged 'control' problem

Another frequent claim made in response to concerns about the need to identify where responsibility lies for the adverse implications of advanced digital technologies is that, because these systems operate more or less autonomously and without direct human intervention and control from the outside, those who develop and implement them cannot fairly be regarded as responsible for their decisions, actions and corresponding consequences. This view was outlined by Matthias[177], who argues that

> the agent can be considered responsible only if he knows the particular facts surrounding his action, and if he is able to freely form a decision to act, and to select one of a suitable set of available alternative actions based on these facts.[178]

But an increasing class of machines, which Matthias refers to as 'autonomous artificial agents', are capable of fulfilling some, often quite narrow, purposes by moving autonomously through some 'space' and acting in it *without human supervision*. That agent can be a software programme that moves through an information space (eg an internet search spider) but it can also have a physical presence (eg a robotic pet) and move through time and space. These agents are deliberately designed to act, and inevitably interact, with other things, people, and social entities (laws, institutions and expectations). At least for those which have a physical presence and can learn from direct interaction in real environments, they can, in return, directly manipulate that same environment and share their environment with humans.

Matthias argues that a 'responsibility gap' arises because, for machine agents of this kind, the human agent who programmed it no longer exerts direct control over the machine agent's behaviour, which is gradually transferred to the machine itself. It would therefore be unjust to hold humans responsible for actions of machines over which they could not have sufficient control.[179] He offers several examples of these kinds of machine agents, including those that rely upon:

---

174 In the legal literature, the term 'victim' or 'potential victim' tends to be used rather than 'moral patient', the latter being more common in the applied philosophical literature.

175 Liu and Zaweiska 2017.

176 Liu 2016.

177 Matthias 2004.

178 Matthias 2004: 175.

179 Matthias's argument has been prominent in shaping the debate, in which the underlying 'choice' theory of moral responsibility upon which his argument rests has not been challenged. Instead, academic responses have either sought to counter his argument via a commitment to methodological and moral individualism such that every action is ultimately attributable to human individuals: whatever role non-human objects played in bringing about a particular outcome, they are ancillary (Hanson 2009: 92). On this view, AI

(a) the operation of *artificial neural networks*: instead of clear and distinct symbolic representation of information and flow control, we have a sometimes very large matrix of synaptic weights, which cannot be directly interpreted. Rather, the knowledge and behaviour stored in a neural network can only be inferred indirectly through experimentation and the application of test patterns after the training of the network is finished;

(b) *reinforcement learning*: usually based on the same neural network concepts, but additionally it lifts the distinction between a training and a production phase. Reinforcement learning systems explore their action space while working in their operational environment, which is their central feature (enabling them to adapt to ever-changing environments) as well as a big drawback concerning their predictability. The information stored in the network cannot be fully checked, even indirectly, because it always changes. Even if we can prove mathematically that the overall performance of such a system will eventually converge to some optimum, there will be *unavoidable errors* on the way to that optimised state. The creator of such a system (who Matthias comments is not really a programmer in the traditional sense) cannot eliminate these errors, for they must be explicitly permitted in order that the system can remain operational and improve its performance;

(c) *genetic programming methods* in which an additional layer of machine-generated code operates between the programmer and the product of programming. Unlike in neural networks, where the designer still defines the operating parameters of the system (the network architecture, the input and output layers, and their interpretation) and at least defines the alphabet used and the semantics of the symbols, the genetic programmer loses even this minimal amount of control, for she creates a machine that programs itself.

At the same time, Matthias observes that autonomous agents deprive the programmer of a spatial link between the programmer and the resulting machine agent. Accordingly, the machine agent acts outside the programmer's observation horizon and might not be able to intervene manually (in the case of a fault or error, which might occur at a much later point in time). Thus, these processes involve the designer of machines increasingly losing control over them, gradually transferring control to the machine itself, in which - according to Matthias – the programmer's role changes 'from *coder* to *creator* of software organisms'. As the influence of the creator of the machine decreases, the influence of the operating environment increases such that the programmer transfers her control over the product to the environment (especially for machines that continue to learn and adapt in their final operating environment). Particularly given that these agents will have to interact with a potentially great variety and number of people (users) and situations, it will typically not be possible for the creator to predict or control the influence of the operating environment. According to Matthias, the net result is that these machines operate *beyond their creators' control*, and may thus cause harm for which we cannot justly hold them responsible. Yet Matthias argues that because we

---

technologies are conceived of as a tool employed by humans so that responsibility for fault will always reside with humans (be they programmers, coders, manufacturers, or developers, users etc): Johnson 2006; Bryson 2010; Sullins 2005. Others have responded by considering AI to signal the instantiation of some moral or legal person of independent ontological status (eg Gunkel 2017) including the ascription of moral agency to computational systems (Dennett 1997; Sullins 2005). However, the weight of academic opinion denies that non-human entities can have moral responsibility in their own right because they lack the mental qualities (and hence cannot meet the epistemic condition) generally accepted as necessary for moral responsibility, which – at least in the philosophical literature, are often expressed in terms of intentionality, the capacity to act voluntarily and an awareness of their actions and anticipated consequences of those actions: Johnson 2006; Kuflick 1999; Sparrow 2007; Asaro 2014 and Hanson 2009: 93.

cannot do without such systems, we must find a way to 'address the responsibility gap in moral practice and legislation.'[180]

(b) Choice-based theories of moral responsibility

Matthias's claim that those who create autonomous machines cannot be 'justly' held responsible for their actions rests on a 'choice-based' account of moral responsibility which has tended to dominate contemporary academic reflection concerning the ethical and moral implications of AI. According to choice-based accounts of moral responsibility, conduct rightly attracts blame when it is at fault, fault being understood in terms of being freely chosen.[181] On this account, an agent (X) is only morally responsible for an unwanted outcome (Y) if X 'caused' Y. To establish that X caused Y, then X must have engaged in conduct for which X can be held causally responsible. Establishing this causal link requires that X *voluntarily chose* to engage in the relevant conduct, even if that conduct turns out to have consequences and effects that X did not intend or want. According to Matthias, because the developers of computational agents which have the capacity to make their own decisions in ways that have not been pre-programmed in advance by human developers, those developers lack the requisite degree of control and therefore are not morally responsible for the decisions of those computational agents or their consequences.[182]

The validity of the claim that the capacity for computational agents to act autonomously breaks the chain of causation between the acts of their developers and the decisions taken by those agents is highly debatable.[183] As a preliminary matter, it is important to recognise that choice theories of moral responsibility are particularly unsuitable as a model for identifying responsibility for *human rights violations*. It is inherent in the nature and concept of rights generally, and human rights in particular, that they protect values of such fundamental importance that any interference with them attracts responsibility per se, *without proof of fault*.[184] Consider again the example of Facebook's removal of the iconic image of the Vietnamese girl in 2016. In circumstances where national legislation imposes legal obligations on both state and non-state actors to respect human rights, Facebook would be regarded as legally responsible for violating the right to freedom of expression, without the need to demonstrate that it had the capacity to control whether the image was removed. In other words, a violation of the right to freedom of expression occurred even if the decision to take-it down had been taken by an automated algorithmic system acting independently without direct human intervention, and even if the human designers of the automated system had not intended or foreseen that the specific image in question might be automatically removed.

3.4 Models for allocating responsibility

Although the model of responsibility that applies to *human rights violations* is widely understood as one of 'strict responsibility', without the need for proof of fault, the allocation of obligations of repair for *tangible harm* to health or property, may be legally distributed in accordance with a variety of responsibility models. Because AI systems might operate in ways that result in both human rights violations and harm to individuals and/or property, and

---

180 Matthias 2004: 183.

181 Wallace 1994, cited by Cane 2002.

182 Matthias 2004. For a recent affirmation, see Gunkel 2017.

183 Ascribing causal responsibility to some action or event is an interpretive act and not a matter of scientific 'truth' *per se*.

184 See UN Special Representative of the Secretary General 2011 (the 'Ruggie Principles')

because the allocation of historic responsibility for harm serves as a guiding function to those involved in the design, development, production and implementation of AI systems by specifying the nature and scope of their obligations, these models are briefly outlined in the following discussion. The variety of legal models that might be applied to allocate and distribute the adverse effects arising from our other-regarding conduct clearly demonstrates that it is a mistake to expect one single model of responsibility to apply fairly to all the different kinds of adverse consequences that might flow from the use of advanced digital technologies. As previously noted, unlike philosophical analysis of responsibility, which tend to focus on agents at the expense of 'victims' and of society, legal models of responsibility[185] are *relational* in the sense that they are concerned not only with the position of individuals whose conduct *attracts* responsibility (i.e. moral agents), but also with the *impact* of that conduct on other individuals and on society more generally.[186] As legal scholar and philosopher Peter Cane has observed

> Responsibility is not just a function of the quality of will manifested in conduct, nor the quality of that conduct. It is also concerned with the interest we all share in security of person and property, and with the way resources and risks are distributed in society. Responsibility is a relational phenomenon.[187]

In other words, legal responsibility emphasises the relationship between moral agents, moral patients and society more generally, rather than focusing exclusively on the conduct of moral agents and whether that conduct justly attracts responsibility. Accordingly, academic analysis of the variety of ways in which national legal systems allocate responsibility for conduct that causes harm or other adverse events (including rights violations that may or may not result in harm) demonstrate how each of these models entails a different balancing of interest between moral agents and moral patients (or 'victims' are they are typically referred to in legal scholarship).[188] This discussion does not, however, seek to evaluate whether current legal approaches adopted within national legal systems adequately allocate responsibility for harm through the application of national civil liability rules, particularly given the capacity of national law to allocate historic responsibility for harms and wrongs by AI systems is yet to be fully tested via litigation.[189] Instead, the following discussion briefly outlines four broad models of responsibility reflected in Anglo-American legal systems, notably (1) intention/culpability-based models (2) risk/negligence-based models (3) strict responsibility and (4) mandatory insurance schemes,[190] as exemplars of different ways in which legal responsibility for risks,

---

185 The concept of 'responsibility' is used much more commonly outside the law to refer to 'human conduct and consequences thereof that trigger such responses' so that we tend to speak of 'moral responsibility' on the one hand and 'legal liability' on the other, with the latter referring primarily to formal institutionalised imposts, sanctions and penalties which are characteristic of law and legal systems but not of morality: Cane 2002: 1-2.

186 Cane 2002: 4-5.

187 Cane 2002: 109.

188 The European Commission is currently undertaking reviewing these issues. See for example European Commission 2018c.

189 Various bodies are working on seeking to evaluate the capacity of national civil liability rules to respond adequately to harm arising from the operation of AI systems. For example, the European Commission intends to produce guidance in mid-2019 to address the way in which the EU Product Liability Directive applies to artificial intelligence, robotics and the Internet of Things: European Commission 2018c.

190 This study identifies various models of responsibility utilised in Anglo-American legal systems for the simple reason that the author of this report has been trained in and is most familiar with the Anglo-American legal system. This should not be taken as an indication that these models are representative of responsibility models reflected in other legal systems, or are in any way superior to models adopted elsewhere.

human rights violations and collective harms might be distributed.[191] They are intended merely as heuristics aimed at highlighting the range of potential models of responsibility that might be used to allocate and distribute threats, risks and harms associated with the use of advanced digital technologies.[192] These sketches therefore selectively describe the what I will refer to as the 'control/conduct condition' and the 'epistemic condition', applicable to each model, rather than providing a complete and detailed account of each model's content and contours. Taken together, they reveal how each model strikes a different balance between our interest, as agents, in freedom of action and our interest, as victims, in rights and interests in security of person and property.[193] It suggests that identifying which (if any) of these models is most appropriate for allocating and distributing the various risks associated with the operation of advanced digital technologies is by no means self-evident[194], but will entail a *social policy choice* concerning how these burdens should be appropriately allocated and distributed.

3.4.1 Intention/culpability-based models

Intention/culpability-based models, which constitute the core model of responsibility that underpins the criminal law, focus primarily on the voluntariness of the agent's conduct. They can be interpreted as requiring the satisfaction of two conditions: firstly, the 'control' condition, demonstrating that the agent was causally responsible for the legally proscribed conduct in so far as the agent had a free and voluntary choice concerning whether so to act, and secondly, the 'epistemic condition,' requiring proof of 'fault,' broadly understood as requiring that the agent had actual knowledge and awareness of the particular facts surrounding the harmful consequences of the agent's conduct, and the agent's action can be understood as based on these facts.[195] It is an intention/culpability-based model of responsibility that underpins the choice-based accounts of moral responsibility that have predominated in philosophically-oriented discussions concerning whether the human developers of autonomous computational agents are morally responsible for the actions of those agents. For the time being at least, because computational agents lack the capacity for subjective knowledge, awareness and intent, these responsibility models cannot be readily applied to computational agents per se because they cannot satisfy the requisite epistemic condition.[196] Intention/culpability-based models can, however, be applied to the human developers or users of such computational agents. The conduct of individuals who

---

191 According to the European Parliament' in its Draft Motion on the Civil Liability Rules on Robotics, 'civil liability for damage caused by robots is a crucial issue which also needs to be analysed and addressed at Union level in order to ensure the same degree of efficiency, transparency and consistency in the implementation of legal certainty throughout the European Union for the benefit of citizens, consumers and businesses alike': European Parliament Committee on Legal Affairs 2016: 16.

192 In Anglo-American legal systems, the distinction between the civil and criminal law is of critical importance. The primary purpose of the criminal law is to impose penalties and punishments on those who engage in criminal conduct, and hence the paradigm of criminal liability focuses primarily on the alleged offender's conduct and mental state. In contrast, the primary purpose of the civil law is to identify and allocate legal obligations of repair on those identified as legally responsible for the relevant harm. Accordingly, responsibility in civil law is two-sided, concerned not only with agent-conduct, but also with the impact of that conduct on others. The operation of the civil and criminal law paradigms of cut-across fault based, negligence-based and strict responsibility models, and the distinction between the civil and criminal law paradigms are not further discussed. For an extensive discussion, see Cane 2002.

193 Cane 2002: 98.

194 Danaher 2016.

195 In Anglo-American law, the mental elements of legal fault criteria are intention, recklessness, knowledge/belief and malice. See Cane 2002: 79.

196 Hildebrandt 2013; Himma 2009; Solum 1991; Gless et al 2016; Andrade et al 2007.

intentionally develop or deploy AI technologies for dangerous or malicious purposes, for example, in order to commit fraud or misappropriate property, would clearly satisfy the requirements for establishing responsibility under an intention/culpability-based model. In these circumstances[197], a prima facie violation of human rights would arise (proof of subjective intent could be shown, but there would be no need to do so because legal responsibility for rights violations is typically 'strict') and would also be likely to generate both responsibility under the criminal law for offences against the person (or property) as well as triggering civil law obligations of repair and restoration.

3.4.2 Risk/Negligence-based models

In Anglo-American law, risk/negligence-based models of legal responsibility for tangible harm form the basis of a general duty to take reasonable care to prevent foreseeable risks of harm. These models of responsibility are conventionally applied to determine whether agents are subject to legal obligations of repair towards those who have suffered harm as a result of an agent's failure to discharge this general duty of care. A 'control condition' similar to that which applies to intention/culpability-based models of responsibility also applies to risk/negligence-based models (with some modification[198]), insofar as it must be shown that the agent caused the relevant damage or injury. However, the epistemic condition applicable to risk/negligence-based models is considerably less demanding than those applicable to intention/culpability-based models. For example, legal liability in negligence under Anglo-American law does not require proof of the agent's accompanying mental state, thereby seeking to strike a fair balance between the interest of agents (in freedom of action) and the interests of victims in safety and security. As legal philosophers have emphasised, in order to hold an agent morally responsible, the agent need not in fact have actual subjective knowledge of the consequences of her behaviour in order to be justly held responsible for it.[199] As John Oberdiek explains, facts matter morally: they are endowed with a normative force that bears upon the permissibility of prospective action but only once they have been reasonably discoverable.[200] In deciding upon a course of action, Oberdiek points out that the ordinary person can be morally expected to take 'reasonable epistemic care': she cannot be expected to know all the facts, but nor can she stick her head in the sand and fall back on her subjective understanding if she has failed to take reasonable care to find out or discover the relevant facts.

Accordingly, whether or not responsibility based on a risk/negligence model can be ascribed to the human developers of computational agents and systems in circumstances where those systems generate decisions or behaviours that cause harm will depend upon whether that harm was a *reasonably foreseeable consequence* of the computational systems' actions and decisions. In Anglo-American negligence law, legal responsibility for causing harm is only ascribed to those who are subject to a legal duty of care. Such a duty arises when, very broadly speaking, there is a reasonably foreseeable risk that an action could harm a proximate person.

---

197 The use of AI technologies in the commission of a crime might appropriately be regarded as an aggravating factor in the commission of a criminal offence: see 6 2002. See also Hallevy 2015.

198 Causation in negligence may be negated by the application of principles of 'remoteness of damage': Horsey and Rackley 2015, chapter 9.

199 Hart 1968.

200 Oberdiek 2017: 57.

Foreseeability therefore operates both to define the kinds of risks for which a person may be legally responsible, and bounds the harms for which they may be liable.[201]

Reasonable foreseeability also plays a role in determining *how* a person is expected to act. The duty of care is discharged if a person acts as an ordinary person exercising reasonable care to avoid foreseeable risks.[202] Hence, reasonable foreseeability operates as the touchstone for determining the relevant 'reference class' for evaluating whether risk-taking activities (such as driving) that may result in tangible harm to others gives rise to a legal duty of care. As Oberdiek observes, this common law standard is a just and appropriate *moral* standard because, in the case of risk-taking activities, it is important that we should be able to hold each other accountable for our respective characterisations of risk. In other words, we must be able to justify that characterisation in a way that withstands moral scrutiny.[203]

Yet in order to identify whether or not it is reasonably foreseeable that any given risky action might ripen into harm, we encounter the so-called 'reference class' problem. As Oberdiek explains:

> the reference class problem is…essentially a problem of redescribability – any particular risk can be infinitely re-described…there is no uniquely correct generative reference class that credible beliefs take as their object.[204]

For example, consider the fatal injury caused by an Uber vehicle which it collided with a woman pushing a bicycle with shopping bags hanging from its handlebars in 2018. The vehicle had been operating in self-driving mode for 19 minutes before it mistook the woman for a car (which it therefore expected would take evasive action), only recognising its mistake and handing back control to the vehicle's human driver seconds before collision, which the human driver was not able to prevent.[205] It seems unlikely that the car's developers could have reasonably foreseen that the vehicle's AI sensing system would mistakenly believe that a woman pushing a bicycle with shopping bags dangling from its handlebars was another vehicle. On the other hand, it seems well within the bounds of reasonable foresight that the car's sensing technologies would fail correctly to classify unusually shaped objects encountered during normal driving conditions, and that errors of this kind might lead to fatal collisions.

At the same time, identifying whether particular events associated with the operation of a particular technological object is 'reasonably foreseeable' will invariably be a product of our experience and exposure to them. In the emerging phases during which a new technology is being rolled out, expectations of their behaviours (and consequences) will be relatively unsettled and unknown.[206] However, as time passes and we become more accustomed to

---

201 When it is an omission or failure to act that causes harm, these criteria manifest in a range of particular ways, such as one has a duty to protect others from risks that arise as a result of one creating a source of danger or because one has assumed responsibility for the other person's interests. See Lunney and Oliphant 2013, chapter 9.

202 Oberdiek 2017: 40.

203 Oberdiek 2017: 48.

204 Oberdiek 2017: 40.

205 Smith 2018.

206 For example, Microsoft' experimental Tay chatbot was designed to learn to converse in human conversational terms by observing and interacting with Twitter users, improving its performance via conversational interaction and in so doing learning how AI programmes can engage with web users in casual conversation. Instead, it quickly learned to parrot a slew of anti-Semitic and other hateful invective that human Twitter users fed the bot, resulting in Microsoft's decision to shut the chatbot down. This kind of response was not in fact anticipated by Tay's developers, yet it could persuasively be argued that responses

their patterns of behaviours and action, those behaviours and actions may become more familiar to developers and therefore more likely to be regarded as reasonably foreseeable. Therefore, developers of those technologies should be held responsible for negligently failing to take steps that would have averted the resulting harm and wrongdoing.[207] Yet even then, this begs the question about our reasonable expectations of the tech industry in making the decision to release emerging technology into real world contexts: we rightly implement demanding governance regimes for new pharmaceuticals, is this not also true of risky advanced digital technologies?[208]

Additional questions arise concerning the minimum standard of care which AI system developers should be responsible for attaining in the design and implementation of autonomous computational systems. Consider again the fatal collision of the Uber vehicle which misclassified a pedestrian wheeling a bicycle as an approaching vehicle. In contemporary discussions, a common refrain is that autonomous cars will be 'safer' than human drivers, thereby suggesting that the relevant comparator is that of a reasonable human driver. But is it appropriate to apply a model of responsibility and the same standard of care that we apply to an ordinary human vehicle driver operating a traditional human-directed car to that of unintended harm resulting from the actions of a self-driving car? Or is it more appropriate to apply the model of responsibility which conventionally applies to product manufacturers to govern the development and operation of self-driving vehicles, which, in contemporary European law systems, is a model of strict responsibility for product defects (discussed below)? In other words, there are important *policy choices* to be made and it is by no means self-evident that the standard of the ordinary human driver provides the most suitable comparator.[209]

3.4.3 Strict responsibility

As this study has already noted, the model of legal responsibility applicable to rights violations (including violations of human rights and fundamental freedoms) is that of strict responsibility, or 'strict legal liability' as it is called in Anglo-American legal parlance. On this model, responsibility attaches to the agent *without* proof of fault, so that legal responsibility for rights violations attaches to those who cause them *regardless* of whether the responsible agent engaged in conduct that breached a legally specified standard of conduct, and regardless of whether the conduct was intended or accompanied by any particular mental state.[210] Of the four varieties of strict liability identified by Cane, three are of direct relevance to this study: right-based, outcome-based and activity-based strict liability.

(a) *right-based* strict liability: arises when legal rights are violated such that any violation of the sphere of protection bounded by the right triggers liability. The classic example is trespass to land: by interfering with the land-owners' right to exclusive dominion over the land, all intrusions without the consent of the land-owner constitute unlawful interference even if it the intruder was in no sense blameworthy. As already noted, violations of human rights fall into this category of cases.

---

of this kind were reasonably foreseeable, given the volume and frequency with which offensive posts are made online via Twitter. See The Guadian 2016.

207 Liu and Zaweiska 2017.

208 Nemitz 2018; Thomas 2017a; Thomas 2017b.

209 Thomas 2017b.

210 Cane 2002: 82.

(b) *outcome based* strict liability: this form of liability rests on the causation of adverse outcomes (i.e. extrinsic consequences) regardless of fault. Contemporary European product liability laws are based on this model which imposes strict liability on manufacturers for defective products that cause harm to natural persons or property.[211] In relation to advanced digital technologies, questions arise concerning what constitutes a relevant 'defect'. Consider again the fatal collision of the Uber vehicle which initially misclassified a pedestrian wheeling a bicycle as another vehicle, handing back control to the human driver as soon as it recognised its error but, however, too late for the human driver to prevent the collision. It could be argued that in these circumstances, the vehicle was not 'defective' in so far as it functioned in precisely the way that its developers intended. On the other hand, if 'defective' is interpreted to mean 'fit for purpose', then the vehicle's failure to correctly classify the pedestrian and take evasive action to avoid the fatal collision could readily be characterised as defective.[212] A similar approach is often applied where the risk of damage is linked to the unpredictability of behaviour of specific risk groups, such as animals. In these cases, liability is attributed to the persons that are considered responsible for supervising the animal, as they are typically regarded as best placed to adopt measures to prevent or reduce the risk of harm.

(c) *activity-based* strict liability arises in connection with a specified activity, such as various 'possession' offences, such as laws which prohibit the possession of guns, knives, illicit substances and so forth. In Anglo-American law, vicarious liability is an important form of activity-based strict liability, where the relevant activity is defined primarily in terms of a relationship with another person, for whose breach of the law the first person is held strictly liable by virtue of that relationship. Vicarious liability applies to the employment relationship, such that an employer will be strictly liable for the unlawful conduct of an employee carried out in the course of his or her duty. Some jurisdictions may adopt a strict liability approach towards those who carry out dangerous activities (e.g. the operator of a nuclear power plant or of an aircraft) or are ultimately responsible for the dangerous activity (e.g. the owner of a vehicle). In such cases, the underlying rationale is that this person has created a risk, and at the same time also derives an economic benefit from this activity.[213]

These various forms of strict liability distribute the risks associated with potentially harmful activity between agents and victims in ways that accord considerable weight to the interests of victims in security of the person and property. In so doing, they recognise that responsibility is not merely a function of the quality of an agent's will manifested in conduct, nor the quality of that conduct: it is also concerned with the interest we all share in security of person and property, and with the way resources and risks are distributed in society, thereby delineating the boundaries of what our responsibilities are.[214]

3.4.4 Mandatory Insurance

Rather than focus on allocating responsibility to potential candidates who can be understood as contributing to the harms and wrongs that might arise from the operation of advanced digital technologies, a society might decide instead to prioritise the need to ensure that all

---

211 See European Union 1985.

212 Strict liability for damage caused by autonomous robots was favoured by the European Parliament's draft motion on Civil Law Rules on Robotics: European Parliament Committee on Legal Affairs 2016.

213 European Commission 2018b.

214 Cane 2002:108-109.

those who are harmed by the operation of these technologies should be financially compensated. This may be achieved by instituting some kind of mandatory insurance scheme (which could be established on a 'no-fault' basis), establishing an insurance fund to which all those harmed by the operation of these technologies could have recourse.[215] Such a scheme might be funded in various ways, including via contributions from the tech industry, but with claims administered by some independent or public authority. One could also simply require firms involved in the value chain through which these advanced digital systems are designed and implemented to take out mandatory liability insurance.[216] While it is beyond the scope of this study to evaluate the desirability of such schemes, they have the benefit of enabling those harmed from the operation of such technologies to seek financial compensation in circumstances where it is difficult to identify precisely which firms ought to be regarded as responsible for the harm, or if the relevant firms have become insolvent. This may become increasingly important as we become more reliant on autonomous intelligent systems which continue to operate long after their human or corporate developers and owners have died or ceased to exist, so that societies may need to develop long-stop institutions such as collective insurance in order to ensure that victims are not systematically left uncompensated.[217] Proposals to confer legal status on intelligent machines in order to facilitate the administration of compensation payments to injured victims have been proposed in this context.[218]

3.5 Responsibility challenges posed by complex and dynamic socio-technical systems

The preceding analysis has proceeded largely on the assumption that, in seeking to assign responsibility for adverse consequences of advanced digital technologies, cause-effect relations can be readily identified. In practice, however, these technologies form an essential component of highly complex and sophisticated socio-technical systems, generating acute challenges in seeking to identify lines of causal, moral and legal responsibility. Three such challenges are briefly outlined in the following discussion: the problem of 'many hands', 'humans in the loop' and the unpredictable effects of complex dynamics that can arise between multiple interacting algorithmic systems.

3.5.1 The problem of 'many hands'

Except in relation to some forms of strict responsibility, the assignment of responsibility for the threats, risks, harms and rights violations (including human rights violations) typically

---

215 The European Parliament's Committee on Legal Affairs recommended a scheme of this kind for harm caused by specific categories of robots, recommending that an obligatory insurance scheme, which could be based on the obligation of the producer to take out insurance for the autonomous robots it produces, should be established, be supplemented by a fund in order to ensure that damages can be compensated for in cases where no insurance cover exist: European Parliament Committee on Legal Affairs 2016 at 20.

216 European Commission 2018b.

217 6:2001 at 429.

218 For example, the European Parliament's Committee for Legal Affairs called on the European Commission to consider creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently : European Parliament Committee on Legal Affairs (2016) at 18. The European Parliament's Policy Department for Citizen's Rights and Constitutional Affairs (the JURI Committee) has emphatically opposed this particular proposal: Nevjans 2016 at 14- 16. These proposals are separate and distinct from academic discussion concerning whether or not robots should be regarded as *moral* agents and entitled to moral rights protection. An examination of the appropriate legal and moral status of AI agents as independent agents in their own right is beyond the scope of this study. See Solum 1991; Koops 2010; Teubner 2006; Teubner 2018.

require an assessment of whether they can be understood as *caused* by the agent. Yet when seeking to assign causal responsibility for some adverse event[219] or effect that could plausibly be regarded as a direct consequence of the operation of any complex socio-technical system (whether or not it utilises AI technologies), one immediately encounters the 'many hands' problem.[220] This problem arises if one adopts an intention/culpability based model of responsibility. First identified in the context of information technology by philosopher of technology, Helen Nissenbaum,[221] the problem of 'many hands' is not unique to computers, digital technology, algorithms or machine learning. Rather, it refers to the fact that a complex array of individuals, organisations, components and processes are involved in the development, deployment and implementation of complex systems, so that when these systems malfunction or otherwise cause harm, it becomes very difficult to identify who is to blame, because such concepts are conventionally understood in terms of individualistic conceptions of responsibility.[222] In other words, causal responsibility is necessarily distributed where complex technological systems are concerned, diluting causation to mere influence.[223]

The 'many hands' problem may be especially acute in seeking to identify the locus of responsibility for harms or wrongs resulting from the development and operation of AI systems, given that they rely on a number of critical components, namely

(a) The *models* that are developed in order to represent the feature space and the optimisation goal which the system is intended to achieve;

(b) *algorithms*, based on these models, which analyse the data to produce outputs which may trigger some kind of 'action' or decision;

(c) The input *data* (which might or might not include personal data) on which those algorithms are trained;

(d) The *human developers* involved in the design of these systems, who must make value-laden decisions about the models, algorithms and data that are used to train the algorithms upon which performance is tested. They include human beings who undertake the task of labelling the data that is used to train the algorithms[224]; and

(e) The larger *socio-technical system and context* in which the algorithmic system is embedded and in which it operates.

Even assuming that we could satisfactorily identify the allocation of moral responsibility for adverse impacts in relation to each of the above components, this is unlikely to ensure that lines of moral responsibility for unintended adverse consequences can be readily identified when they are dynamically combined within a complex integrated system. These challenges are compounded by the fact that digital products and services are open to software extensions, updates and patches after they have been implemented. Any change to the software of the system may affect the behaviour of the entire system or of individual components, extending their functionality, and these may change the system's operational risk

---

219 The relevant adverse event might be some systemic risks/harm, individual harm or an individual human rights violation not necessarily entailing material loss or damage or harm to collective interests.

220 Thompson 1980.

221 Nissenbaum 1996.

222 Thompson 1980.

223 Liu and Zaweiska 2017.

224 Zalnieriute et al 2019.

profile, including its capacity to operate in ways that might cause harm or violate human rights.[225]

In responding to these challenges, it may be helpful to bear three considerations in mind. Firstly, issues relating to the allocation of legal responsibility for harm arising from activities involving multiple parties are not new, and many legal systems have therefore developed a relatively sophisticated set of principles and procedures for determining liability where multiple potential defendants are involved.[226] As the European Commission has recently observed, identifying the distribution of liability for redress amongst multiple actors involved in the value chain through which emerging digital technologies operate may not be relevant for the purposes of ensuring that victims obtain compensation for damage suffered, although resolving such questions is likely to be important from an overall policy standpoint in order to provide legal certainty to those involved in the production and implementation of these technologies.[227] Secondly, and relatedly, the law's ability to devise practical responses despite the apparent intractability of the many hands problem can be at least partly attributed to the greater emphasis which it places on the legitimate interests of the moral patient in security of the person, rather than the almost exclusive focus on the moral agent that is reflected in choice theories of moral responsibility (and upon which the 'many hands' problem rests). Thirdly, because the focus of this report is on responsibility for *human rights violations* arising from the development and implementation of advanced digital technologies, rather than on responsibility for *harm*, it is particularly important to ensure that we have effective and legitimate mechanisms that will operate to *prevent and forestall* human rights violations, particularly given that many human rights violations associated with the operation of advanced digital technologies may not result in tangible harm to individual health or property. The need for a preventative approach is especially important given the speed and scale at which these technologies now operate. The resulting cumulative and aggregate effects of human rights violations caused by the operation of AI systems could seriously erode the social foundations necessary for moral and democratic orders that are essential preconditions for human rights to exist at all, suggesting that existing approaches to human rights protection may need to be reinvigorated in a networked, data-driven age.[228]

3.5.2 Human-Computer Interaction

Not only are many individuals, firms and other organisations involved in the development and implementation of advanced digital technologies, but these technologies are often intended to operate in ways that involve retaining active human involvement.[229] This points to serious challenges associated with identifying the appropriate distribution of authority and

---

225 Thomas 2015.

226 See for example models of shared responsibility for liability of on-line hosting platforms: eg De Streel, Buiten and Peitz 2018; Helberger et al 2018.

227 European Commission 2018b: 20-21

228 See section 3.8 below.

229 Human oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or a human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system: see EU High Level Expert Group 2019a: 16

responsibility between humans and machines, given the complex interaction between them. In particular, many tasks previously performed by humans are now undertaken by machines, yet humans are invariably involved at various points throughout the chain of development, testing, implementation and operation. As the Royal Academy of Engineering has observed:

> There will always be humans in the chain, but it is unclear in the case of injury which human in the chain bears responsibility – the designer, manufacturer, programmer, or user.[230]

The interaction between humans and machines within complex and dynamic socio-technical systems generate especially challenging questions concerning the appropriate role of humans in supervising their operation. One recurring theme has been a concern that, in order to ensure that increasingly complex socio-technical systems always operate in the service of humanity, systems should always be designed so that they can be shut down by a human operator. Yet, as the Royal Academy of Engineering has again observed:

> It might be thought that there is always need for human intervention, but sometimes autonomous systems are needed where humans might make bad choices as a result of panic – especially in stressful situations – and therefore the human override would be problematic. Human operators are not always right nor do they always have the best intentions. Could autonomous systems be trusted more than human operators in some situations?[231]

On the other hand, even if humans are retained 'in the loop' with the aim of supervising computational systems, individuals placed in these positions may be understandably reluctant to intervene. Over a decade ago, Johnson and Powers[232] commented:

> In the case of future automated air traffic control…there will be a difficult question about whether and when human air traffic controllers should intervene in the computer-control of aircraft…Those humans who formerly held the role responsibility for the duties will either by replaced by caretakers of the technology, or will themselves become caretakers. A concern in this environment is that the humans assigned to interact with these 'automatic' systems may perceive intervention morally risky. It is better, they may reason, to let the computer system act and for humans to stay out of the way. To intervene in the behaviour of automated computer systems is to call into doubt the wisdom of the system designers and the 'expertise' of the system itself. At the same time, a person who chooses to intervene in the system brings the heavy weight of moral responsibility upon him or herself, and hence human controllers will have some incentive to let the automaticity of the computer system go unchallenged. This is a flight from responsibility on the part of humans, and it shows how responsibility has been re-assigned, in some sense, to the computer system.[233]

Yet, as we increasingly rely on the expanding range of services and systems that automation makes possible, particularly as our digital technologies grow ever more powerful and sophisticated, continued insistence on placing a human in the loop to act in a supervisory capacity risks turning humans placed in the loop into 'moral crumple zones' – largely totemic humans whose central role becomes soaking up fault, even if they only had partial control of

---

230 Royal Academy of Engineering 2009: 2.

231 Royal Academy of Engineering 2009: 3.

232 Johnson and Powers 2005: 106.

233 On the question of how far humans can responsibly transfer decision-making functionality to computer without at the same time reserving oversight-responsibility to humans, see Kuflik 1999.

the system, and who are vulnerable to being scapegoated by tech firms and organisations seeking to avoid responsibility for unintended adverse consequences.[234] As Elish and Twang's study of aviation autopilot litigation highlights, modern aircraft are now largely controlled by software, yet pilots in cockpits remain legally responsible for the aircraft's operation. Yet our cultural perceptions tend to display 'automation bias', elevating the reliability and infallibility of automated technology whilst blaming humans for error (see Box 2).[235]

Box 2: Automation bias and the responsibility of humans in the loop

The collision of a Tesla car in semi-automated mode exemplifies the tendency to blame the proximate humans in the loop for unintended adverse consequences, rather than the surrounding socio-technical system in which the human is embedded.

A semi-automated Tesla collided with a truck in May 2016 due to the vehicle's autopilot's failure to detect the truck. The official investigation following the collision revealed that although the autopilot functioned as designed, it did not detect the truck. The human failed to respond, with the investigation concluding that the driver had over-relied on automation and the monitoring steering wheel torque, which were not effective methods for ensuring driver engagement.

The authority undertaking the investigation concluded that the crash was not the result of any specific defect in the autopilot system, so that Tesla was not responsible for the accident. Because Tesla had provided an adequate warning to customers, indicating that the autopilot system must be operated under the supervision of the human driver, and that the driver's hands should remain on the wheel and their eyes on the road, responsibility lay with the human driver. In addition, Tesla's Terms of Services included provisions that referred to the semi-autonomous nature of the autopilot, stating that the driver was to take over the control of the car in 4 seconds if the driver noticed problematic vehicle behaviour.

Source: European Commission, Staff Working Document, 'Liability for Emerging Digital Technologies' (April 2018) 14-15.

3.5.3 Unpredictable, dynamic interactions between complex socio-technical systems

Even more intractable challenges arise in seeking to identify, anticipate and prevent adverse events arise from the *interactions* between complex, algorithm-driven socio-technical systems that can occur at a speed and scale that was simply not possible in a pre-digital, pre-networked

---

234 Elish 2016.

235 Elish points to the tragedy of Air France Flight 447 in 2009 (which crashed into the Atlantic Ocean, killing all 228 people on board) as a classic example of the positioning of individual pilots as moral crumple zones. The flight had flown into a storm en route from Brazil to France, resulting in ice crystals forming on the plane's pitot tubes, part of the avionics system that measures air speed. The frozen pitot tubes sent faulty data to the autopilot which, in turn, reacted in precisely the way in which it was designed to react in the absence of data: it automatically disengaged, transferring control of the aircraft back to the human pilots. The pilots were caught by surprise, overwhelmed by an avalanche of information – flashing lights, loud warning signals, and confusing instrument readings, with the official French report concluding that they 'lost cognitive control of the situation', with a series of errors and incorrect manoeuvres by the pilots resulting in the fatal crash. Elish observes that news coverage of the accident report emphasised the pilots' errors, but failed to draw attention to the fact that many of these errors were at least partly due to the automation, by changing the very kind of control that can be exercised by a human operator, and by creating opportunities for new kinds of error: Elish 2016 ibid.

age. The so-called 'flash crash' that occurred in 2010, during which the stock market went into freefall for five minutes before correcting itself, for no apparent reason, provides a vivid illustration.[236] While individual AI agents, that have the capacity to learn from their environment and to iteratively improve their performance, might be subject to mathematical verification and testing, identifying how multiple different algorithms might *interact* with other algorithmic agents in a complex and dynamic ecosystem generates risks of unpredictable, and potentially dangerous, outcomes. In other words, these interactions generate risks that we have barely begun to grasp.[237] The challenge of devising solutions that will enable us reliably to predict, model and take action to prevent unwanted and potentially catastrophic outcomes arising from the interaction between dynamic and complex socio-technical systems generates a new and increasingly urgent frontier for computational research. Leading computer scientists Shadbolt and Hampson warn of the dangers of "hyper-complex and super-fast systems" generating considerable new risks, and for which:

> Our response needs to be vigilant, intelligent and inventive. So long as we are, we will remain in control of the machines, and benefit greatly from them. We need to develop policy frameworks for this. Beyond the dangers, a world of opportunity arises.[238]

3.6 State responsibility for ensuring effective protection of human rights

One of the most significant concerns about the emergence of algorithmic systems has been the increasing power of Big Tech firms, including concerns about the radical power asymmetry between these firms and the individuals who are subject to them.[239] Accordingly, it is in the hands of these firms that the power to deploy algorithmic systems overwhelmingly resides. Yet, the obligation to protect human rights in the international domain law lies primarily on nation states, given that human rights protection is primarily intended to operate vertically, to protect individuals against unjustified interference by the state. However, it is well established in ECHR jurisprudence that the rights protected by the Convention ground positive substantive obligations requiring member states to take action in order to secure to those within their jurisdiction the rights protected by the Convention.[240] Accordingly, states are obliged under the ECHR to introduce national legislation and other policies necessary to ensure that ECHR rights are duly respected, including protection against interference by *others* (including tech firms) who may therefore be subject to binding legal duties to respect human rights.[241] It is these enforceable legal obligations, grounded in the Convention's protection of human rights, including the right to an effective remedy, that offers solid foundations for imposing legally enforceable and effective mechanisms to ensure accountability for human rights violations, well beyond those that the contemporary rhetoric of 'AI ethics' in the form of voluntary self-regulation by the tech industry can realistically be expected to deliver.[242]

---

236 Akansu 2017.

237 Smith 2018.

238 Shadbolt and Hampson 2018.

239 Ibid, Schwab et al 2018; The Economist 2018b.

240 Rainey, Wicks and Ovey 2014: 102.

241 The scope and extent of the required protection will depend upon the particular right in question. Ibid.

242 The Pilot Judgement Procedure of the European Court of Human Rights provides an institutional mechanism through which states can be directed to adopt individual remedial measures in their domestic legal orders in order to bring to an end violations found by the Court, supervised by the Committee of Ministers. See Glas 2014.

The discussion of various models for allocating historic responsibility outlined in the section 3.2 draws largely on Anglo-American legal approaches introduced via legislation (and adjudicated on by courts) or developed by courts in their interpretation and application of the common law in determining legal liability for harm or other wrongdoing. One significant drawback associated with reliance upon judicial remedies to redress these concerns is that they are better suited to remediating substantial harms suffered by the few, as opposed to less significant harms suffered by the many. The difficulties of seeking redress via the courts are magnified in the AI space by the challenge of detecting the harm and determining and proving causation, to say nothing of the serious practical obstacles and disincentives faced by individuals in invoking the judicial process.[243] At the same time, the capacity of AI systems in a globally networked environment to generate collective action problems has already been highlighted, underlining the need for, and importance of, properly resourced national enforcement equipped with adequate enforcement powers authorities and which may also suggest that accessible and convenient collective complaint mechanisms may be necessary to ensure that enforcement action is taken in relation to human rights violations resulting from the operation of AI systems. At the same time, it is important to recognise that, in addition to conventional legal mechanisms of redress via the courts, there are many *other institutional governance mechanisms* that could help secure responsible human rights-compliant development and implementation of advanced digital technologies. The following section therefore provides a brief outline of other possible institutional governance mechanisms (beyond 'voluntary' self-regulatory initiatives currently emerging) that may serve to enhance both prospective and retrospective responsibility for the threats, risks, harms and wrongs arising from the operation of advanced digital technologies. It briefly outlines several possible mechanisms and governance institutions that might have an invaluable role to play in securing accountability for human rights violations that could complement existing legal mechanisms.

3.7 Non-judicial mechanisms for enforcing responsibility for advanced digital technologies

Although regulatory governance mechanisms can be classified in many different ways, three features are worth highlighting for the purposes of this study. Firstly, we can distinguish between mechanisms which operate on an *ex ante* basis, which provide oversight and evaluation of an object, process or system *before* it has been implemented into real world settings and are therefore primarily concerned with securing prospective responsibility. On the other hand, there are *ex post* mechanisms that operate during or after implementation has occurred and are therefore primarily concerned with securing historic responsibility. As this study has already emphasised, both dimensions of responsibility must be attended to in order to secure the responsible development and implementation of AI systems. Yet because this study is primarily concerned with the human rights implications of these technologies, the need for effective and legitimate mechanisms that will *prevent and forestall* human rights violations is of considerable importance, particularly given the speed and scale at which AI systems can now operate, combined with a culture of 'move fast and break things' that characterises the operational strategy of leading tech firms. This strategy consists of forging ahead with rapid technological innovation without attending carefully to their potential risks in advance, preferring to deal with any adverse 'blow-back' after the event by which time it may not be practically possible to unwind or roll-back the technological innovations that have already been brought to market.[244] Secondly, it is important to attend to the *legal enforceability* of regulatory governance institutions and mechanisms in order to identify whether, and to what extent, they are to be regarded as optional mechanisms which the tech

243 Mantelero 2018: 55.

244 Taplin 2018; Vaidhyanathan 2011.

industry has the freedom selectively to adopt or ignore altogether, or whether they are legally mandated and for which substantial legal sanctions attach for non-compliance.[245] Thirdly, although regulatory governance mechanisms have conventionally taken the form of social institutions, in the present context, the role of *technical protection mechanisms*, which rely upon a modality of control sometimes referred to as 'regulation by design'[246] may be equally (if not more) important. It is to these that this study now turns.

3.7.1 Technical protection mechanisms

One of the most promising fields of research that has flourished in response to growing awareness of the ethical and legal concerns raised by the use of AI technologies, can be found in the *technical* responses that have emerged with the aim of seeking to 'hard-wire' particular values into the design and operation of algorithmic techniques that are incorporated into AI systems.[247] One of the features often associated with some of these 'design-based' regulatory governance mechanisms is their capacity to operate in real time, rather than on an ex ante or ex post basis.[248] Although early work in the field utilising technical measures to secure the protection of particular interests and values through the use of ICT focused primarily on technological solutions to the protection of IP rights[249], parallel work also began to take place in the field of data privacy, which became known as 'privacy by design' or 'data protection by design'. This work recognised that technology could be applied in the service of interests and values that it concurrently threatened, seeking to improve the bite of legal norms on IP rights and data privacy by seeking to build the norms into information systems architecture.[250] In addition to the work on 'privacy engineering', more recent research in machine learning and software engineering can be understood as building on this approach, seeking to secure what may be called 'human rights protection by design' and include the following:

(a) Explainable AI (XAI): Advances in machine learning techniques, including those relying on neural networks (NN), are often used to aid human decision making,[251] yet their logic is not easily explainable (i.e., when they opt for a particular choice, we do not know why they do so) or readily interpretable (i.e., they cannot explain or present outcomes in ways humans can understand). There is growing recognition of the need to ensure that outputs generated by AI systems can be rendered intelligible to users[252] and this has opened up a significant field of computational research in 'explainable AI' (XAI).[253]

(b) Fairness, Accountability and Transparency in Machine Learning (FATML): Similarly, a growing community of ML researchers have directed their attention towards developing techniques to identify and overcome problems of 'digital discrimination'[254] referring to bias and discrimination arising from the use of data mining and other machine learning

---

245 Nemitz 2018.
246 Yeung 2015.
247 Ibid.
248 Ibid.
249 These were originally referred to as 'Electronnic Copyright Management Systems (ECMS) and later referred to as Digital Rights Management Systems (DRMS).
250 Bygrave 2017.
251 See for example, Doshi-Velez, Ge, & Kohane, 2013; Carton et al 2016.
252 Weller 2017; Yeung and Weller 2018b.
253 See for example Samek et al 2017; Wierzynski 2018.
254 Barocas and Selbst 2016; Criado and Such 2019**;** Zliobaite 2015.

techniques (known as 'discrimination-aware' or 'fairness-aware' techniques for machine learning).[255]

3.7.2 Regulatory governance instruments and techniques

More conventional, social and organisational forms of regulatory governance instruments have also emerged in response to recognition that AI technologies might be utilised in ways that could undermine important values, including those explicitly concerned with ensuring that these technological systems operate in ways that respect human rights. Two are briefly discussed here: human rights impact assessment and algorithmic auditing techniques. [256]

(a) Algorithmic/Human rights impact assessment: Various scholars and organisations have proposed various forms of 'algorithmic impact assessment' that are, in effect, proposed risk-assessment models that are to be applied by those seeking to procure or deploy algorithmic systems in order to identify the human rights, ethical and social implications of their proposed systems, and to take steps to ameliorate those concerns in the design and operation of algorithmic systems prior to implementation. While various general impact assessment models have been proposed, a number of domain-specific models have also been proposed.[257]

These risk assessment models vary widely in terms of their:

- *Criteria of assessment*: while EU data protection law now mandates the use of 'Data Protection Impact Assessments (DPIAs)[258]' in certain circumstances, building on pre-existing approaches to 'Privacy Impact Assessment' they are largely focused on the evaluation of impacts upon data quality and security. Other models, such as 'Human Rights Impact Assessment[259]' are concerned with evaluating the impact of a proposed system on human rights more generally.[260]

- *Party undertaking the assessment:* some proposed models are intended to be applied by the data controller (eg DPIAs) while others propose that the assessment be undertaken by an external third party or accreditation body, which is the approach reflected in the UN

---

255 See in particular the annual event organised by FATML (http://www.fatml.org/) and resources listed at http://www.fatml.org/resources/relevant-scholarship.

256 See UN General Assembly 2018 which endorses both techniques.

257 For example, in relation to the use of algorithmic decision-making systems by the pubic sector, see AI Now Institute: 2018. That Report outlines a framework for public sector entities in the US to use in carrying out 'algorithmic impact assessments', prior to purchasing or deploying an automated decision. In relation to criminal justice risk assessment, see Selbst 2018 and Oswald et al 2018. In relation to the human rights risks for internet registries, see ARTICLE 19: 2017.

258 Article 35 of the EU General Data Protection Regulation (GDPR) requires the preparation of data protection impact assessments where data processing is likely to result in a 'high risk' to the rights and freedoms of natural persons.

259 Various models of human rights impact assessment can be understood as more specific forms of 'human rights due diligence', growing out of the UN Guiding Principles on Business and Human Rights, which uses the term 'due diligence' as 'the essential first step towarad identifying, mitigating and redressing the adverse human rights impacts of AI' per Rasso et al 2018: 53. See also Toronto Declaration on Machine Learning 2018.

260 Mantelero 2018.

Guiding Principles on Business and Human Rights in relation to 'human rights due diligence'[261].

- *Mandatory or voluntary adoption*: Some algorithmic-human rights impact assessment proposals are intended to be adopted on a voluntary basis, so that it is up to the data controller to choose whether or not to undertake the assessment and what, if any, steps to take in light of that assessment.[262] Others, such as the DPIA, are mandated by law if certain threshold conditions are satisfied.[263]

- *Scale of evaluation*: While Human Rights Impact Assessment is concerned with scrutinising a wide range of business operations to assess their conformity with human rights standards, other forms of impact assessment, such as the DPIA or PIA, are much narrower in their scale of evaluation, focusing on a single data processing activity.

  Impact assessment techniques can be valuable in focusing attention on the various ways in which a proposed activity may risk interfering with human rights, in ways that might otherwise be overlooked or ignored. Yet, in order for impact assessment approaches to provide real and substantive protection, it will be necessary to develop a clear and rigorous methodological approach that firms and other organisations are willing to adopt consistently and in ways that reflect a genuine commitment to identifying human rights risks, rather than merely regarding them as a bureaucratic burden resulting in 'ritual' displays of formal compliance without any genuine concern to respect human rights.[264]

(b) Algorithmic auditing: Unlike impact assessment approaches which are intended to take place *before* system implementation, algorithmic auditing techniques are aimed at testing and evaluating algorithmic systems once they are in operation. Algorithmic auditing is emerging as a field of applied technical research, that draws upon a suite of emerging research tools and techniques for detecting, investigating and diagnosing unwanted adverse effects of algorithmic systems.[265] It has been proposed that techniques of this kind might be formalised and institutionalised within a legally mandated regulatory governance framework, through which algorithmic systems (or at least those algorithmic systems that are regarded as 'high risk' systems in terms of the seriousness and scale of the consequences in the event of failure or unintended adverse effects) are subject to periodic review and oversight by an external authority staffed by suitably qualified technical specialists. For example, Cukier and Mayer-Schonenberg suggest that a new group of professionals are needed ('algorithmists') to take on this role, which may constitute a profession akin to that of law, medicine, accounting and engineering and who can be relied upon to undertake the task of algorithmic auditing either as independent and external algorithmists to monitor algorithms from the outside, or by 'internal' algorithmists employed by organisations to monitor those developed and deployed by the organisation, which can then be subjected to external review.[266]

---

261 Raso et al 2018: 53.

262 Mantelero 2018.

263 See above n. 241; Mantelero 2018 ; Edwards and Veale 2017.

264 Power 1997.

265 Desai and Kroll 2017. Sandvig et al 2014. See for example the resources available at Auditing Algorithms.

266 Such an approach would resemble the governance of conventional financial auditing, in which the accounting systems within organisations are subject to both internal auditors employed in house, and also from external auditors, who then legally obliged to review those accounts and certify their veracity and validity: Mayer-Schonberger and Cukier 2013: 180. See also Crawford and Schultz 2014; Citron 2008.

3.7.3 Standard setting, monitoring and enforcement

The techniques and approaches described above have significant potential as instruments through which prospective and historic responsibility for systems that rely upon advanced digital technologies might be secured. Yet in order for this potential to be realised, we must also attend to the legal and institutional governance frameworks in which they are embedded. For example, the various strands of technical research referred to at section 3.7.2 have considerable potential to facilitate prospective responsibility for digital technologies, providing welcome recognition by the technical community that digital systems are not 'neutral' but are imbued with values and might act in ways that are not consistent with human rights. Not only is it important that this work is nurtured and supported, but it is also important that it emerges from interdisciplinary engagement between the technical community and those from law, the humanities and the social sciences, in order to elaborate more fully how human rights values can be translated into technical mechanisms of protection, and how a human rights approach responds to the problem of value conflict. It is equally important that we attend to the *legal status* of these techniques. Although the tech industry has been keen to adopt technical responses to ethical problems, merely placing 'blind faith' in industry solutions risks becoming merely another form of 'ethics washing'.[267] In other words, unless these technical approaches are themselves backed by law and subject to transparent evaluation by and oversight by a competent independent authority to ensure their validity and operation, they may not provide effective human rights protection. As regulatory governance scholarship has emphasised, it is vital that all three elements of the regulatory governance process are attended to: the setting of standards, information gathering and monitoring of activity that is required to comply with those standards, and enforcement action and sanctions for non-compliance.[268] Effective and legitimate regulatory governance requires both stakeholder participation in the setting of the relevant standards and a properly resourced, independent authority equipped with adequate powers systematically to gather information, to investigate non-compliance and to sanction violations.[269] If we are to have confidence that technological protection mechanisms intended to ensure that human rights values are respected during the operation of digital processes, then we must have robust mechanisms of oversight that can investigate and verify that they do in fact so operate. Hence technical standards themselves should be developed independently (and ideally through a participatory process in which all affected stakeholders can be involved) and subject to external scrutiny and examination, and that compliance with those standards can and will be scrutinised by an external body who has the power to impose (or seek to ensure the imposition of) sanctions for violation. In other words, without meaningful independent oversight, these mechanisms are unlikely to provide the foundations for securing meaningful human rights accountability. Various national and local governments are increasingly recognising the need for more formal, institutionalised and systematic consideration and evaluation of algorithmic systems, reflected in the various task-forces and public authorities commissioned to provide review and/or oversight of data-driven socio-technical systems.[270]

3.8 Reinvigorating human rights discourse in a networked digital age

As we enter a new globally networked digital age, the need to protect human rights and the underlying value commitments upon which they rest, is of paramount importance. This

---

267 Greene et al 2019.

268 Morgan and Yeung 2007; Lodge and Wegrich 2014.

269 Nemitz 2018.

270 For a summary of national initiatives across Europe, see Access Now 2018.

prompts consideration of whether our existing conceptions of human rights and the mechanisms through which they are enforced are fit for purpose in this new socio-technical landscape. The powerful networked digital technologies that have emerged in recent years make possible practices and actions that were previously impossible and thereby create novel threats, risks and forms of wrong-doing, provoking reflection on whether additional new human rights and regimes of institutional governance are required to ensure that those risks can be meaningfully addressed in practice.[271] Although the basic structure and institutional framework for human rights protection, which is well-established and universally recognised, can reasonably be expected to develop effective responses to many of the threats and challenges wrought by the rising power of digital automation and machine intelligence, there are several reasons why our existing rights discourse and enforcement mechanisms may require reinvigoration if they are to provide effective protection. Firstly, many of the rights conferred upon data subjects are difficult to assert in practice, largely due to the opacity of many of the socio-technical systems in which these technologies are embedded. Secondly, our understanding of the scope and content of existing rights were developed in a pre-networked age. So conceived, these rights might fail to provide comprehensive protection against the full range of threats and risks to individuals which these technologies may give rise to, particularly in relation to illegitimate attempts to deceive and manipulate individuals that so-called 'persuasive technologies may enable (see above) and problems of discrimination (see above). For example, although the rights to data protection confer upon the data subject a right to insist upon human intervention, to express her view or to contest a fully automated decision that has 'profound effects' on her, these rights do not apply to partially automated decisions. Nor do they necessarily ensure that, in practice, an affected individual can readily detect whether she has been treated unequally vis-à-vis others, and if so, whether such differential treatment amounted to discrimination and was thus prima facie unlawful. Thirdly, and perhaps most importantly when considering the adequacy of existing human rights and fundamental freedoms to address the new risks associated with new digital technologies, is the data subject's freedom to *waive* some of these rights by consenting to specific practices that would otherwise constitute a rights-violation, thereby forgoing the protections these rights provide.[272] For example, if individuals were to rely only on Article 8 to protect the rights and interests implicated in the provision of data-driven services, there is a significant risk that these rights would be too readily waived by individual right-holders in a networked age built upon a 'free services' business model: thus, in return for 'free' access to digital services and the efficiency and convenience they offer, individuals willingly exchange their personal data.[273] In contrast, the core data protection principles upon which contemporary European data protection regimes (including modernised Convention 108) rest (and reflected in the jurisprudence of the European Court on Human Rights under Article 8), include mandatory obligations imposed on data controllers that cannot be waived by individual right-holders,

---

271 Brownsword, Scotford and Yeung 2017.

272 The extent to which the European Court On Human Rights is willing to recognise the possibility of individuals waiving their ECHR rights, and the conditions required for an effective waiver, is likely to depend upon the right in question and the specific context in which a claimed waiver is alleged to arise. For example, *Scoppola* v. *Italy (No. 2)*, 17 September 2009, no. 10249/03, para. 135, the Court stated "Neither the letter nor the spirit of Article 6 prevents a person from waiving them of his own free will, either expressly or tacitly. However, such a waiver must, if it is to be effective for Convention purposes, be established in an unequivocal manner and be attended by minimum safeguards commensurate with its importance [...]. In addition, it must not run counter to any important public interest [...]." In relation to private law and contractual relationships between non-state actors, the Court is likely to consider the issue of waiver in terms of the positive duty of states to take reasonable measures to protect individuals from infringement of Convention rights by other private persons, including the obligation to ensure (through legal regulation or other measures) that the relevant rights are 'practical and effective' in their exercise.

273 Solove 2012.

including the principles of lawfulness of the processing, of purpose specification and data minimisation, thereby offering more systematic protection of the core underlying values and collective interests which these regimes ultimately seek to protect.

But quite apart from these potential weaknesses, the individualised orientation of contemporary conceptions of human rights and existing mechanisms for their enforcement, may fail to give due attention to the threats which these technologies may pose to collective goods, particularly the need to preserve and nourish the underlying socio-technical foundations that make it possible for moral agency and human rights to have space to operate. Leading philosopher of law and technology, Mireille Hildebrandt, expresses these concerns in terms of the technical conditions that are assumed to exist in order for the law (and contemporary understandings of the rule of law) to fulfil its functions.[274] Yet within 'smart' environments that operate by continuously collecting digital data from the material world in order to infer, predict and therefore anticipate future behaviour of things, people and systems, these technical conditions are both supplanted and augmented, thereby altering the very possibility for the exercise of what we currently understand as thought, choice and reason because smart technologies operate continuously and immanently, and because they are designed to learn, producing outcomes that their designers did not specify.[275]

North American jurist Julie Cohen develops Hildebrandt's insights, drawing on both legal scholarship and a growing body of work in the sociology of science referred to as Science and Technology Studies (or 'STS').[276] Cohen argues that to ensure that human rights can be operationalised in an era of smart environments, we must 'take affordances seriously', otherwise our rights will be ineffective. According to affordance theory, the design of our technological objects and environments condition and constrain the possibilities for action, including the range of actions and responses which the design of the object 'affords' to the user. Thus, once we recognise that smart digital technologies continually, immanently and pre-emptively mediate our beliefs and choices, then our legal discourse about human rights (including privacy) can be understood as incomplete. Cohen therefore persuasively argues that this requires more than merely extending rights discourse. Rather, it will require us to *reconceive of rights in new ways*, as well as developing a *different vernacular for rights discourse* – one that recognises the central role of sociotechnical configurations in affording and constraining the freedoms and capabilities that people in fact enjoy.[277] In particular, our rights discourse has operated on a set of often unexamined assumptions about the built environment's properties, about both constraint (such as the physical impossibility of universal surveillance) and lack of constraint (such as the open-ended possibilities for spaces people use to gather and assemble, for various purposes including democratic protest). But advances in networked digital technologies are challenging these assumptions, and we are only now learning that the relevant constraints and affordances include not only those affecting our physical space, but also the affordances that govern the flow of data and information, and that these have direct impacts for our rights and freedoms. We therefore need to expand our frame of rights discourse to encompass our socio-technical architecture, in which rights can be conceived in terms of affordances as a practical matter in ways that 'speak with effective force to new kinds of material and operational considerations.'[278]

---

274 Hildebrandt 2015.

275 Cohen 2017 at 3 citing Hildebrandt 2015 at 88-102.

276 Cohen 2017.

277 Cohen 2017: 7.

278 Cohen 2017: 9.

In other words, the inability of rights to provide a comprehensive response to the threats posed by AI technologies is more deeply rooted in the inherent limitations of rights-based approaches effectively to address systematic harms that are experienced primarily at a collective, societal level, rather than at the level of the individual right-holder. For example, the introduction of a new 'right to meaningful human contact'[279] has its attractions, but it might not be effective in addressing the concerns about systematic, societal dehumanisation which lies at the foundation of many of the anxieties expressed about our increasing reliance on computational technologies. In other words, it is the aggregate and cumulative effects of these technologies over time, and at scale, that may systematically threaten the socio-technical foundations which the very notion of human rights presupposes and in which they are rooted.[280]

Because smart digital technologies are 'radically different in kind' from other kinds of technologies, the societal challenge is to contend with their difference and power.[281] By focusing on the architectural implications of these technologies, our attention is drawn to a perspective that Cohen describes as 'inherently communal'. It highlights the responsibility of states, and our collective responsibility as a moral community, to attend to the socio-technical foundations of moral and democratic freedom, and the way in which the aggregate, cumulative impact of the adverse social concerns referred to above could fundamentally undermine the 'moral and democratic commons'[282] and without which human rights and fundamental freedoms cannot, in practice, be realised or asserted. These social foundations must, at minimum, ensure that conditions necessary for moral agency and responsibility are present and secure, for in their absence, there is no freedom, and human rights have no meaning.[283] Yet we lack institutional mechanisms for monitoring the health of the socio-technical foundations in which our human rights and democratic freedom are anchored, and this may require us to develop both a new 'vocabulary' of rights, and institutional mechanisms for ensuring the health and sustainability of these foundations to secure meaningful human rights protection in a new hyper-connected digital age.[284]

3.9 Summary

This section has highlighted the importance of ensuring that responsibility for the actual and potential adverse consequences associated with the development and operation of advanced digital technologies is allocated prospectively and retrospectively. The fair and effective allocation of responsibility for these threats, risks and adverse impacts is vital, not only to protect human rights and safeguard the welfare of individuals, groups and society at large , but also, and even more fundamentally, to ensure that our society remains a moral community. Yet attributing responsibility for the adverse risks and effects of our increasingly powerful and sophisticated digital technologies generates considerable challenges, owing to the fact that there are a great many individuals and organisations involved in their development and implementation, and because they may operate in unexpected ways.

Welcome recognition of the need to take seriously responsibility for the risks and other adverse effects of advanced digital technologies can be found in the proliferation of voluntary

279 Discussed above at section 2.2.2.
280 Yeung 2011.
281 Hildebrandt 2015; Cohen 2017.
282 Yeung 2011; Yeung 2017b.
283 Brownsword 2005.
284 Yeung 2011.

initiatives through which the tech firms and the tech industry have promulgated codes of good ethical practice, which they publicly proclaim they will aspire to meet. Yet because these voluntary self-regulatory initiatives lack any institutional mechanisms for meaningful public participation in the setting of the relevant standards, nor any external enforcement and sanctioning mechanisms, they do not constitute legitimate and effective safeguards.

Although the capacity of advanced digital systems to operate more or less autonomously has been claimed to distance their developers from responsibility for their operation, this claim rests on a very particular and narrow conception of moral responsibility. We have seen that a range of responsibility models might be available for allocating responsibility for the adverse impacts of AI systems, noting that in relation to human rights infringements, responsibility is appropriately assigned on a strict basis, without proof of fault. As states bear the primary duty for ensuring effective protection of human rights, this grounds a legal obligation to introduce national legislative frameworks that give rise to legal duties and obligations on non-state actors. In addition, the fundamental value of human rights is of such strength and importance that they are increasingly recognised as grounding horizontal effects on non-state actors, including tech developers.[285] While judicial remedies constitute an important avenue through which those adversely affected by the operation of AI technologies might seek redress, we have also identified a range of other governance instruments (including technical protection mechanisms) that could be utilised to secure meaningful and effective accountability and which warrant further consideration.

Yet although there are various governance mechanisms (described above) that, if backed by law, can help to secure meaningful human rights protection, they are – in and of themselves – unlikely to provide adequate and comprehensive protection. In particular, our advanced networked digital technologies are now of such power and sophistication that they can be understood as 'radically different in kind' from other kinds of technologies, particularly given their profound implications for our collective and shared technical, social, democratic and moral architecture of our societies. We must therefore reinvigorate our existing human rights discourse and instruments in ways that foreground our collective responsibility to attend to the socio-technical foundations of moral and democratic freedom, and the way in which the aggregate, cumulative impact of the adverse social concerns referred to above could fundamentally undermine the 'moral and democratic commons'and without which human rights and fundamental freedoms cannot, in practice, be realised or asserted.

---

[285] For the private sector, this has most comprehensively been developed by UN Special Rapporteur Ruggie who 'codified' the corporate social responsibility to *respect* human rights and act accordingly even in countries where national legislation does not demand that.

Chapter 4. Conclusion

Advances in techniques now referred to as artificial intelligence are likely to continue to develop and grow in power and sophistication in the foreseeable future. Relatively recent success in AI, combined with the global and interconnected data infrastructure that has emerged over time, have enabled the proliferation of digital services and systems. These have already delivered very considerable benefits, particularly in terms of the enhanced efficiency and convenience which they offer across a wide range of social domains and activities, although access to these remains largely the province of inhabitants of wealthy industrialised nations. They bring with them extraordinary promise, with the potential to deliver very substantial improvements to our individual and collective well-being, including the potential to enhance our capacity to exercise and enjoy our human rights and freedoms. Yet, there are also legitimate and rising public anxieties about their adverse societal consequences, including their potential to undermine human rights protection which, as this study has highlighted, could threaten to destabilise the very foundations upon which our moral agency ultimately rests. This study has therefore sought to examine the implications of advanced digital technologies (including AI) on the concept of responsibility from a human rights perspective. It has identified a series of 'responsibility relevant' properties of these technologies, outlining a range of adverse impacts which these technologies may generate, and has sought to identify how responsibility for preventing, managing and mitigating those impacts (including the risk of human rights violations) may be allocated and distributed.

This study has shown that any legitimate and effective response to the threats, risks, harms and rights violations potentially posed by advanced digital technologies is likely to require a focus on the consequences for individuals and society which attends to, and can ensure that, both *prospective responsibility* aimed at preventing and mitigating the threats and risks associated with these technologies, and *historic responsibility*, to ensure that if they ripen into harm and/or rights violations, responsibility for those consequences is duly and justly assigned. Only then can we have confidence that sustained and systematic effort will be made to prevent harms and wrongs from occurring, and that if they do occur, then the underlying activities will be brought to an end, and that effective and legitimate institutional mechanisms for ensuring appropriate reparation, repair, and prevention of further harm are in place. It will necessitate a focus on both those involved in the development, deployment and implementation of these technologies, individual users and the collective interests affected by them, and on the role of states in ensuring the conditions for safeguarding individuals subject to their jurisdiction against risks and ensuring that human rights are adequately protected.

Four findings of this study are worth highlighting:

1. It is particularly important to ensure that we have effective and legitimate mechanisms that will operate to *prevent and forestall* human rights violations, particularly given that many human rights violations associated with the operation of advanced digital technologies may not result in tangible harm. The need for a preventative approach is especially important given the speed and scale at which these technologies can operate, and the real risk that such violations may erode the collective socio-technical foundations that are essential for freedom, democracy and human rights to exist at all. This has several implications. Firstly, it suggests that states have an important responsibility to ensure that they attend to the larger socio-technical environment in which human rights are anchored. Secondly, stronger collective complaints mechanisms may be needed to ameliorate the collective action problem that individuals may encounter in responding to rights violations generated by the operation of AI systems. Thirdly, our existing conceptions of human rights may need

to be reinvigorated in a networked, data-driven age in order to account for the way in which these technologies may reconfigure our socio-technical environments, and the threats which they may pose to collective goods and values.

2. The model of legal responsibility that applies to *human rights violations* is widely understood as one of 'strict responsibility', without the need for proof of fault. In contrast, the allocation of obligations of repair for *tangible harm* may be legally distributed in accordance with a variety of responsibility models (briefly outlined in Section 3.4 above). This variety of potential legal models that could be applied to allocate and distribute the adverse effects arising from our other-regarding conduct clearly demonstrates that it is a mistake to expect one single model of legal responsibility to fairly apply to all the different kinds of adverse consequences that might flow from the use of advanced digital technologies. Legal models of responsibility emphasise the relationship between moral agents, moral patients and society more generally, unlike much applied philosophical analysis of responsibility for AI systems, which has tended to focus on the conduct of moral agents and whether that conduct justly attracts responsibility agents at the expense of moral patients ('victims') and of society. These various legal models of responsibility strike a different balance between our interest as agents in freedom of action, and our interest as victims in rights and interests in security of person and property. Identifying which (if any) of these models is most appropriate for allocating and distributing the various risks associated with the operation of advanced digital technologies is by no means self-evident, but will entail a deliberate *social policy choice* concerning how these risks should be appropriately allocated and distributed. In democratic societies that espouse a commitment to human rights, the state bears a critical responsibility for ensuring that these policy choices are made in a transparent, democratic manner which ensures that the policy ultimately adopted will effectively safeguard human rights.

3. Various strands of technical research have considerable potential to help secure prospective and historic responsibility for advanced digital technologies through the development of techniques that may enable both effective technical protection mechanisms and meaningful 'algorithmic auditing'. This research should be nurtured and supported, and needs to be developed through interdisciplinary engagement between the technical community and those from law, the humanities and the social sciences, in order to elaborate more fully how human rights norms can be translated into technical mechanisms of protection, and how a human rights approach responds to the problem of value conflict.

4. Taking human rights seriously in a hyperconnected digital age will require that have effective and legitimate governance mechanisms, instruments and institutions are in place to monitor and oversee the development, implementation and operation of our complex socio-technical systems. Some suggestions for how we might take forward the need to ensure that we have governance mechanisms and institutions that have the capacity to do this are set out in Appendix A. Voluntary initiatives by the tech industry via the promulgation of so-called 'ethical' standards of conduct which they publicly claim they will seek to honour constitute welcome recognition by the tech industry that the technologies which they develop may produce adverse effects for which they bear some responsibility. They do not, however, provide adequate and robust human rights protection. At minimum, responsible development and implementation of AI requires both democratic participation in the setting of the relevant standards and the existence of properly resourced, independent authorities

> equipped with adequate powers systematically to gather information, to investigate non-compliance and to sanction violations. In particular, if we are to have confidence that technological protection mechanisms intended to ensure that human rights values are respected during the operation of digital processes, then we must have robust independent mechanisms of external oversight that can investigate and verify that they do in fact so operate, otherwise they are unlikely to provide the foundations for securing meaningful AI accountability. In this respect, it is the obligation of states to ensure that these governance mechanisms are established and implemented in ways that will ensure the protection of human rights.

If we are serious in our commitment to protect and promote human rights in a hyperconnected digital age, then we cannot allow the power of our advanced digital technologies and systems, and those who develop and implement them, to be accrued and exercised without responsibility. The fundamental principle of reciprocity applies: those who deploy and reap the benefits of these advanced digital technologies (including AI) in the provision of services (from which they derive profit) must be responsible for their adverse consequences. It is therefore of vital importance that nations committed to protect human rights uphold a commitment to ensure that those who wield digital power (including the power derived from accumulating masses of digital data) are held responsible for their consequences. It follows from the obligation of states to protect human rights that they have a duty to introduce into national law, governance arrangements that will ensure that both prospective and historic responsibility for the adverse risks, harms and rights violations arising from the operation of advanced digital technologies are duly allocated.

Appendix A

This appendix identifies a range of measures and institutional mechanisms that might warrant further consideration and research in order to help ensure that human rights are protected in an age of advanced networked digital technologies. They are not intended as recommendations, but merely to invite further reflection and discussion.

Prospective responsibility

Consider offering additional funding to support and encourage interdisciplinary research aimed at developing techniques, mechanisms and standards that can help ensure that prospective responsibilities for preventing and mitigating risks of harm or wrongs arising from the operation of advanced digital technologies are duly assigned.

Consider measures to encourage states and interstate cooperation to work towards developing legally supported institutional governance mechanisms to facilitate the protection of human rights against threats and risks posed by advanced digital technologies. These might include:

a. Legal requirements to undertake 'human rights impact analysis' (incorporating algorithmic impact analysis) prior to deployment of advanced digital technologies, including a publicly available statement identifying how potential interferences with human rights and value conflicts are resolved in system architecture and operation;

b. Develop, in conjunction with a wide range of stakeholders, a code of best practice for preparing human rights impact analysis for advanced digital technologies.

c. Clarify the scope and content of legal obligations of all those involved in the development of digital services (including software developers), particularly obligations that bear directly upon human rights protection;

d. Consider the need to subject developers and providers to legal obligations to engage in, and demonstrate adequate verification and testing of, complex computational systems that may have a direct and substantial impact on human rights, both prior to release and at periodic intervals following implementation in real-world environments;

e. Encourage the use of technical protection mechanisms (such as 'human rights by design', fairness-aware data mining techniques, and explainable AI), identifying how they can serve a valuable role in ensuring human rights adherence. Consider the need to provide legal support for these techniques, including by subjecting them to external oversight and review in order to provide a greater level of assurance that these mechanisms operate in fact in ways that are human rights compliant;

f. Encourage further research into the development of techniques and standards that support responsible, human-rights compliant innovation in digital tech industry (including modelling, data provenance and quality, algorithmic auditing, validation, verification and testing).

g. Consider establishing a professional accreditation scheme for appropriately qualified technical experts trained in algorithmic auditing techniques as a class of professionals who are subject to fiduciary duties of loyalty and good faith in verifying and certifying the design and operation of algorithms.

h. Develop a methodological framework and set of metrics for systematically identifying, and evaluating the magnitude and seriousness of, potential threats and risks to individual rights (including the threats they pose to the socio-technical foundations in which human rights and fundamental freedoms are anchored) posed by proposed or potential AI applications.

i. Consider whether AI applications which pose threats that are judged to be so serious and disproportionate in their human rights impact that they should be prohibited unless they are subjected to prior public consultation and approval from an appropriately constituted independent supervisory authority. A framework of this kind might include a class of AI applications that should be prohibited outright because they pose unacceptable grave and potentially catastrophic threats to human rights and fundamental freedoms.[286]

Historic responsibility

Consider supporting the development of guidance and techniques that can help ensure that historic responsibility is duly assigned for individual and collective harms or rights violations resulting from the operation of advanced digital technologies. This may include encouraging states and intergovernmental cooperation towards developing legally supported institutional governance mechanisms that might include:

a. Member state action to review and assess whether national legal systems will operate to ensure that responsibility for harm caused by advanced digital technologies can be duly allocated, identifying any potential gaps which may need to be addressed via legislative reform;

b. Consider the need to develop standard-setting instruments to clarify and locate default historic responsibility for the harms and wrongs to those involved in the design, developers, deployment, ownership and provision of digital systems. This could include legal liability to make reparation to those harmed or wronged by the operation of these services, including an obligation to compensate and introduce measures to avoid future occurrence. In developing a suitable instrument, consideration might be given to the desirability of some kind of 'due diligence' defence in certain clearly and narrowly defined circumstances, leading to a reduction in the extent of the developer's legal responsibility for harm or wrongdoing;

c. Support further research into the appropriate distribution and allocation of authority between humans in the loop of complex computational systems, in light of the acknowledged problem of 'automation bias' and tendency to allocate responsibility to individual humans in the loop, rather than on those who develop and implement the socio-technical system in which the human is embedded;

[286] See also EU High Level Group on Artificial Intelligence (2019a).

d. Consider the desirability of mandating a compulsory insurance regime for the digital tech industry, including whether to establish a national insurance scheme, funded by digital tech industry, to ensure that victims are not left uncompensated;

e. Support the development of further capacity to establish new (and extend the capacity of existing) governance institutions that can meaningfully and rigorously investigate and enforce prospective and historic responsibilities of digital service developers and providers.

f. Consider the desirability of introducing collective complaints mechanisms and whether to liberalise standing rules in order to overcome the problem of collective action which may arise when a large number of individuals may be vulnerable to rights infringement but are unlikely to be sufficiently motivated to take action even though their cumulative effect may be very substantial. To this end, consider whether the collective complaints procedure adopted to enhance the effectiveness, speed and impact of the implementation of the European Social Charter provides a suitable model.

g. Review adequate resourcing and powers of investigation, sanction and remedies for public enforcers. This may include the need to develop and build technical expertise and competence in machine learning and other software development and evaluation techniques within the public sector.

<u>Reconfiguring human rights discourse in a networked digital age</u>

Consider ways in which existing human rights protection and discourse may need to develop in order to ensure the effective protection of human rights in a globally connected digital age, recognising the need to attend to the socio-technical foundations that form the basis of the rule of law and of moral community. This might include:

a. Consider the desirability for a new Convention on Human Rights in a Networked Digital Age which would, at minimum, recognise that both prospective and historic responsibility for risks, harms and rights violations must be fully allocated and distributed;
b. Consider the need for formal recognition within such a Convention (or other similar multilateral instrument) of the role of independent institutional mechanisms to safeguard against the collective risks which these technologies pose to the social foundations of democratic orders in which human rights are anchored;
c. Consider whether new collective decision-making and monitoring mechanisms may be necessary or desirable in order to track and evaluate the aggregate and cumulative effects of these technologies on human rights across member states. To this end, consider the need or desirability of establishing a 'global observatory' to undertake this monitoring and reporting function on a systematic basis;
d. Apply a precautionary approach in cases where interacting algorithmic systems have the capacity to cause catastrophic harm which could not reasonably have been foreseen by any individual digital service provider; consider the prohibition of particular kinds of algorithmic applications with the potential for causing catastrophic harms ; consider the need for systematic monitoring structures and expert institutions in order to prevent such applications from being developed and deployed.

References

6, P. (2001) 'Ethics, regulation and the new artificial intelligence, part I: accountability and power'. *Information, Communication & Society* 4(2):199-229.

6, P. (2001) 'Ethics, regulation and the new artificial intelligence, part II: autonomy and liability'. *Information, Communication & Society* 4(3): 406-434.

6, P. (2002) 'Who wants privacy protection, and what do they want?' *Journal of Consumer Behaviour: An International Research Review*. 2(1): 80-100.

Access Now (2018) *Mapping Regulatory Proposals for Artificial Intelligence in Europe*. Available at https://www.accessnow.org/mapping-artificial-intelligence-strategies-in-europe/ (Accessed 7.11.18)

AI Now (2017) *AI Now 2017 Report*. Available at https://ainowinstitute.org/AI_Now_2017_Report.pdf (Accessed 31.10.2018).

Akansu, A. N. (2017). 'The flash crash: a review.' *Journal of Capital Markets Studies* 1(1): 89-100.

Amnesty International (2017) *Artificial Intelligence for Good*. Available at https://www.amnesty.org/en/latest/news/2017/06/artificial-intelligence-for-good/(Accessed 2.11.2018).

Andrade, F., Novais, P., Machado, J. and Neves, J. (2007) 'Contracting agents: legal personality and representation' *Artificial Intelligence and Law*. 15(4): 357-373.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) 'Machine bias'. *ProPublica,* 23 May. Available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed 5.11.2018).

ARTICLE 19, The Danish Institute for Human Rights and the Dutch Internet Domain-registry (2017) *Sample ccTLD Human Rights Impact Assessment Tool*. Available at https://www.article19.org/wp-content/uploads/2017/12/Sample-ccTLD-HRIA-Dec-2017.pdf (Accessed 2.11.2018).

Asaro, P.M. (2014) 'A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics' in *Robot Ethics: The ethical and social implications of robotics.* Edited by P. Lin, K. Abney & G.A. Bekey. MIT Press.

Auditing Algorithms: Adding Accountability to Automated Authority. Available at http://auditingalgorithms.science/ (Accessed 5.11.2018).

Australian Human Rights Commission (2018) *Human Rights and Technology Issues Paper.* July. Available at https://tech.humanrights.gov.au/sites/default/files/2018-07/Human%20Rights%20and%20Technology%20Issues%20Paper%20FINAL.pdf (Accessed 5.11.18).

Barocas, S. and Selbst, A.D. (2016) 'Big data's disparate impact.' *Cal. L. Rev*. 104: 671.

Barr, S. (2018) 'Computer-Generated Instagram Account Astounds Internet'. *The Independent*. 1 March. Available at https://www.independent.co.uk/life-style/fashion/instagram-model-computer-generated-shudu-gram-internet-cameron-james-a8234816.html (Accessed 7.11.18)

Bennett Moses, L. and de Koker, L. (2017). 'Open Secrets: Balancing Operational Secrecy and Transparency in the Collection and Use of Data by National Security and Law Enforcement Agencies' *Melbourne University Law Review* 41(2): 530.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Bovens, M. (2007) 'New forms of accountability and EU-governance' *Comparative European Politics* 5(1): 104-120.

boyd, d., and Crawford, K. (2012) 'Critical Questions for Big Data'. *Information, Communication and Society* 15(5):662-79.

Brownsword, R. (2005) 'Code, control, and choice: why East is East and West is West.' *Legal Studies 25*(1): 1-21.

Brownsword, R., Scotford, E., and Yeung., K. (eds.) (2017). *Oxford Handbook on Law, Regulation and Technology*. Oxford: Oxford University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. and Anderson, H. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.'* Available at *arXiv preprint arXiv:1802.07228* (Accessed 5.11.2018).

Bublitz, J.C. (2013) 'My mind is mine!? Cognitive liberty as a legal concept' in *Cognitive Enhancement: An Interdisciplinary Perspective,* edited by E. Hildt, and A.G Franke. Dordrecht: Springer at 233-264.

Burrell, J. (2016) 'How the machine 'thinks': Understanding opacity in machine learning algorithms.' *Big Data & Society* 3(1):1-12.

Bryson, J. J. and A. Theodorou (2018) 'How Society can Maintain Human-Centric Artificial Intelligence'. In M. Toivonen-Noro and E. Saari (eds.) *Human-Centered Digitalization and Services*.

Bryson, J.J. (2010) 'Robots should be slaves' in *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues,* edited by Y Wilks. Amsterdam: John Benjamins Publishing at 63-74.

Bygrave, L.A. (2017) 'Hard-wiring Privacy' in Brownsword, R., Scotford, E., and Yeung., K. (eds.) (2017). *Oxford Handbook on Law, Regulation and Technology*. Oxford: Oxford University Press.

Cane, P. (2002) *Responsibility in Law and Morality*. Oxford: Hart Publishing.

Carton, S., Helsby, J., Joseph, K., Mahmud, A., Park, Y., Walsh, J., Cody, C., Patterson, C.P.T., Haynes, L. and Ghani, R., (2016) 'Identifying police officers at risk of adverse events.' In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, at 67-76. Available at https://dl.acm.org/citation.cfm?id=2939698 (accessed 5.11.18).

Cath, C. (2018) 'Governing artificial intelligence: ethical, legal and technical opportunities and challenges' 376 *Phil Trans A*: *Mathematical, Physical and Engineering Sciences*. https://doi.org/10.1098/rsta.2018.0080

Chen, A. (2014) 'The Labourers Who Keep Dick Pics and Beheadings Out of Your Facebook News Feed', *Wired*, 23 October. Available at https://www.wired.com/2014/10/content-moderation/ (Accessed 5.11.18).

Chesney, B. and D. Citron (2019) 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.' *California Law Review* 107: forthcoming.

Citron, D. K. (2008) 'Technological Due Process.' *Washington University Law Review* 85*:* 1249-1313.

Cohen, J. E. (2017). 'Affording Fundamental Rights.' *Critical Analysis of Law.* 4(1): 76-90.

Conn, A. (2017) 'Research for Beneficial Artificial Intelligence' *Future of Life Institute.* Available at: https://futureoflife.org/2017/12/27/research-for-beneficial-artificial-intelligence/?cn-reloaded=1&cn-reloaded=1 (Accessed 5.11.18)

Cowley, J. (2018) 'Beijing subway to install facial recognition as fears grow of China surveillance powers.' 19 June. *The Telegraph*. Available at https://www.telegraph.co.uk/news/2018/06/19/beijing-subway-install-facial-recognition-fears-grow-china-surveillance/ (Accessed 5.11.18)

Council of Europe. Recommendation CM/Rec(2018)2 of the Committee of Ministers to member states on the roles and responsibilities of internet intermediaries. Adopted on 7 March 2018. Available at https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14 (Accessed 7.11.18)

Council of Europe, Parliamentary Assembly (2017) Committee on Culture, Science, Education and Media. *Technological Convergence, Artificial Intelligence and Human Rights*. 10 April. Doc 14288. Available at http://semantic-pace.net/tools/pdf.aspx?doc=aHR0cDovL2Fzc2VtYmx5LmNvZS5pbnQvbncveG1sL1hSZWYvWDJILURXLWV4dHIuYXNwP2ZpbGVpZD0yMzUzMSZsYW5nPUVO&xsl=aHR0cDovL3NlbWFudGljcGFjZS5uZXQvWHNsdC9QZGYvWFJlZi1XRC1BVC1YTUwyUERGLnhzbA==&xsltparams=ZmlsZWlkPTIzNTMx. Accessed 7.11.18.

Crawford, K. and Schultz, J. (2014) 'Big data and due process: Toward a framework to redress predictive privacy harms.' *Boston College Law Review* 55:93.

Danaher, J. (2016) 'Robots, law and the retribution gap.' *Ethics and Information Technology* 18(4): 299-309.

Datta, A., Sen, S. and Zick, Y. (2016) Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*, 598-617. IEEE.

Davidow, B. (2014). 'Welcome to Algorithmic Prison - The use of Big Data to to profile citizens is subtly, silently constraining freedom'. *The Atlantic*. 20 February.

Dennett, D.C., (1997) 'When HAL kills, who's to blame' in *HAL's Legacy: 2001's Computer as Dream and Reality,* edited by D.G. Stork. MIT Press.

Desai, D.R. and Kroll, J. (2017) 'Trust but Verify: A Guide to Algorithms and the Law'. *Harvard Journal of Law & Technology* 31:1-64

De Streel, A., Buiten, M.& Peitz, M. (2018) 'Liability of online hosting platforms: should exceptionalism end?' *Centre on Regulation in Europe Report.* Available at: http://www.cerre.eu/sites/cerre/files/180912_CERRE_LiabilityPlatforms_Final_0.pdf (Accessed 5.11.18)

Dietrich, W., Mendoza, C and Brennnan, T. (2016) 'Compass risk scales: Demonstrating accuracy, equity and predictive parity'. Northpointe.

Donahoe, E. (2016) 'So Software Has Eaten the World: What Does it Mean for Human Rights, Security and Governance?' 18 March, *Just Security*. Available at https://www.justsecurity.org/30046/software-eaten-world-human-rights-security-governance/ (Accessed 5.11.18).

Doshi-Velez, F., Ge, Y. and Kohane, I. (2014) 'Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis.' *Pediatrics 133*(1): e54-e63.

Draper, N. A. and J. Turow (2017) 'Audience Constructions, Reputations and Emerging Media Technologies: New Issues of Legal and Social Policy' in *The Oxford Handbook on Law, Regulation and Technology*, edited by R. Brownsword, E. Scotford & K. Yeung. Oxford: Oxford University Press.

The Economist (2017) 'Imitating people's speech patterns could bring trouble'. Available at https://www.economist.com/science-and-technology/2017/04/20/imitating-peoples-speech-patterns-precisely-could-bring-trouble (Accessed 7.11.18)

The Economist (2018a) 'Images aren't everything: AI, radiology and the future of work'. Available at https://www.economist.com/leaders/2018/06/07/ai-radiology-and-the-future-of-work (Accessed 6.11.18)

The Economist (2018b). 'The techlash against Amazon, Facebook and Google - and what they can do'. Available at https://www.economist.com/briefing/2018/01/20/the-techlash-against-amazon-facebook-and-google-and-what-they-can-do (Accessed 6.11.18)

Edwards, L. and Veale, M. (2017) 'Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking for.' *Duke L. & Tech. Rev.* 16.

Ekbia, H. and B. Nardi (2014) 'Heteromation and its (dis)contents: The invisible division of labor between human and machine. *First Monday* 19(6). Available at https://firstmonday.org/article/view/5331/4090#author (Accessed 7.11.18)

Elish, M.C. (2016): 'Letting Autopilots Off the Hook: Why do we blame humans when automation fails?' 16 June. Available at: http://www.slate.com/articles/technology/future_tense/2016/06/why_do_blame_humans_when_automation_fails.html (Accessed 5.11.18)

Engineering and Physical Sciences Research Council (EPSRC): https://epsrc.ukri.org/ (Accessed 6.11.18)

Eschelman, A., (2016) 'Moral responsibility', *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). Edited by Edward N. Zalta. Available at https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/ (Accessed 6.11.18)

European Commission (2018a) *Communication on Artificial Intelligence.* 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.

European Commission (2018b) Consumer market study on online market segmentation through personalised pricing/offers in the European Union, 19 July. Available at https://ec.europa.eu/info/publications/consumer-market-study-online-market-segmentation-through-personalised-pricing-offers-european-union_en (Accessed 3 May 2019).

European Commission (2018c) *Liability for emerging digital technologies.* European Commission Staff Working Document. COM (2018) 237 final. Available at https://ec.europa.eu/digital-single-market/en/news/european-commission-staff-working-document-liability-emerging-digital-technologies (Accessed 5.11.2018)

European Commission (2018d) Evaluation of Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability. COM(2018) 246 final. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018SC0157&from=EN (Accessed 7.11.18)

European Commission (2019a) High Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI.*

European Commission (2019b) High Level Expert Group on AI, *A Definition of AI: Main Capabilities and Disciplines*.

European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe. COM (2018) 237 final. Available at https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe (Accessed 5.11.2018)

European Group on Ethics in Science and New Technologies (EGE)(2018). *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*. Available at https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf (Accessed 6.11.18)

European Convention on the Protection of Human Rights and Fundamental Freedoms (ECHR)

European Parliament Committee on Legal Affairs (2017) *Report with Recommendations to the Commission on Civil Law Rules on Robotics.* Rapporteur: M. Delaveux. 2015/2103 (INL)). Available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2017-0005+0+DOC+PDF+V0//EN (accessed 5.11.2018).

European Political Strategy Centre (2018) 'The age of artificial intelligence: Towards a European Strategy for Human-Centric Machines'. Available at https://ec.europa.eu/epsc/sites/epsc/files/epsc_strategicnote_ai.pdf (Accessed 6.11.2018)

European Union (1985) Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products (OJ L 210, 7.8.1985, 29-33).

Executive Office of the President. (2016) Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. Available at: https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf (Accessed 6.1118).

Farr, C. (2016) 'If You Want Life Insurance, Think Twice Before Getting A Genetic Test'. 2 July. Available at https://www.fastcompany.com/3055710/if-you-want-life-insurance-think-twice-before-getting-genetic-testing (Accessed 7.11.18)

Ferguson, A.G. (2016) 'Policing predictive policing.' *Wash. UL Rev.* 94: 1109.

Ferraris, V., Bosco, F. and D'Angelo, E.,(2013) 'The impact of profiling on fundamental rights'. Available at: SSRN: https://ssrn.com/abstract=2366753 or http://dx.doi.org/10.2139/ssrn.2366753 (Accessed 6.11.2018).

Financial Times (2018) FT Series , *The AI arms race.* Available at https://www.ft.com/content/21eb5996-89a3-11e8-bf9e-8771d5404543

ForbrukerRadet. Norweigan Consumer Council (2018) *Deceived by Design*. 27 June. Available at https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf (Accessed 7.11.18).

Galligan, D.G. (1997) *Due Process and Fair Procedures*. Clarendon Press: Oxford.

Galligan, D.G. (2006) *Law in Modern Society*. OUP: Oxford.

Gandy, O.H. (1993) *The panoptic sort: a political economy of personal information.* Westview.

Gardner, J. (2003) 'The Mark of Responsibility.' *Oxford Journal of Legal Studies.* 23(2): 157-171.

Gardner, J. (2008) *Introduction to H.L.A. Hart, Punishment and Responsibility: Essays in the Philosophy of Law: Second Edition.* OUP Oxford.

The Guardian (2016) 'Microsoft 'deeply sorry' for racist and sexist tweets by AI chatbot'. Available at: https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot (Accessed 6.11.18)

Gilliker, P (2000) 'A "new" head of damages: damages for mental distress in the English law of torts'. *Legal Studies* 20: 19-41.

Girardin, F. and Blat, J. (2010) The co-evolution of taxi drivers and their in-car navigation systems. *Pervasive and Mobile Computing*. 6(4): 424-434.

Glas, L.R. 'The Functioning of The Pilot-Judgment Procedure Of The European Court Of Human Rights In Practice' (2016) *Netherlands Quarterly of Human Rights* 34(1): 41.

Gorton, W.A. (2016) 'Manipulating Citizens: How Political Campaigns' Use of Behavioral Social Science Harms Democracy' *New Political Science*, *38*(1), pp.61-80.

Greene, D., Hoffman, A.L and Stark, L., 'Better, Nicer, Clearer and Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning' (2019) Hawaii International Conference on System Sciences, DOI: 10.24251/HICSS.2019.258. Available at http://dmgreene.net/wp-content/uploads/2018/09/Greene-Hoffman-Stark-Better-Nicer-Clearer-Fairer-HICSS-Final-Submission.pdf (Accessed 6 May 2019)

Gunkel, D.J. (2017) 'Mind the gap: responsible robotics and the problem of responsibility', *Ethics and Information Technology*, pp.1-14.

Hagendorf, T. (2019) 'The Ethics of AI Ethics: An Evaluation of Guidelines'. Available at https://arxiv.org/abs/1903.03425 (Accessed 6 May 2019).

Hall, W. and Pesenti, J (2017) *Growing the Artificial Intelligence Industry in the UK*. Available at https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk (Accessed 7.11.18)

Hallevy, G. (2015) *Liability for crimes involving artificial intelligence systems*. Springer International Publishing.

Hanson, F.A. (2009) 'Beyond the skin bag: on the moral responsibility of extended agencies', *Ethics and information technology*, *11*(1), pp.91-99.

Hart, H.L.A., ((1968) 2008) *Punishment and responsibility: Essays in the philosophy of law*. Oxford University Press: Oxford.

Helberger, N., Pierson, J., and Poell, T. (2018) Governing online platforms: From contested to cooperative responsibility. *The Information Society* 34(1): 1-14, DOI:10.1080/01972243.2017.1391913

Hildebrandt, M. and Gutwirth, S., (2008) Profiling the European Citizen. Springer: Netherlands.

Hildebrandt, M. (2013) 'Criminal Law and Technology in a Data-Driven Society' in M.D. Dubber and T Hornle (eds), *Oxford Handbook of Criminal Law*. Oxford: Oxford University Press 174-197.

Hildebrandt, M., (2015) *Smart technologies and the end (s) of law: novel entanglements of law and technology*. Edward Elgar Publishing.

Hildebrandt, M. (2016) 'Data-gestuurde intelligentie in het strafrecht' In: E.M.L. Moerel, J.E.J. Prins, M.Hildebrandt, T.F.E. Tjong Tjin Tai, G-J. Zwenne & A.H.J. Schmidt (eds.), *Homo Digitalis*. Nederlandse : Wolters Kluwer.

Himma, K.E. (2009) 'Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent have to be a Moral Agent.' *Ethics and Information Technology* 11(1):24.

Horsey, K and Rackley, E (2014) *Tort Law*. 4th ed. Oxford University Press: Oxford.

Hutson, M (2018) 'Lip-reading artificial intelligence could help the deaf—or spies.' 31 July. *Science*. doi:10.1126/science.aau9601. Available at http://www.sciencemag.org/news/2018/07/lip-reading-artificial-intelligence-could-help-deaf-or-spies (Accessed 6.11.18)

Huxley, A. (1932) *Brave New World.* Chatto & Windus.

IEEE, Global Initiative for Ethical Considerations in AI and Autonomous Systems (2017). Available at https://standards.ieee.org/industry-connections/ec/autonomous-systems.html (Accessed 7.11.18)

Irani, L. (2015) 'Difference and Dependence among Digital Workers: The Case of Amazon Mechanical Turk'. *The South Atlantic* Quarterly 114(1): 225-234.

Jasanoff, S. (2016) *The ethics of invention: technology and the human future*. WW Norton & Company.

Johnson, D.G. (2006) 'Computer systems: Moral entities but not moral agents.' *Ethics and information technology*. 8(4): 195-204.

Johnson, D.G. and Powers, T.M. (2005) 'Computer systems and responsibility: A normative look at technological complexity', *Ethics and information technology*, *7*(2), pp.99-107

Kaminski, M.E. and Witnov, S. (2014) 'The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech.' *U. Rich. L. Rev.*, 49: 465.

Keen, A (2018) *How to Fix the Future.* Atlantic Books: London.

Kitchin, R (2014) *The Data Revolution*. Sage: Los Angeles.

Korff, D. and Browne, I. (2013) 'The use of the Internet & related services, private life & data protection: trends, technologies, threats and implications', Council of Europe, T-PD(2013)07. Available at at SSRN: https://ssrn.com/abstract=2356797 (Accessed 6.11.18)

Koops, B.J., Hildebrandt, M and Jaquet-Chiffelle, D-O. (2010) 'Bridging the Accountability Gap: Rights for New Entities in the Information Society?' *Minnesota Journal of Law, Science and Technology* 497.

Kosinski, M., Stillwell, D & Graepel, T. (2013) 'Private traits and attributes are predictable from digital records of human behaviours.' *Proceedings of the National Academy of Science* 110: 5802-5805.

Kramer, A.D., Guillory, J.E. and Hancock, J.T. (2015) 'Experimental evidence of massive-scale emotional contagion through social networks.' *Proceedings of the National Academy of Sciences*, 8788–8790.

Kuflik, A. (1999) 'Computers in control: Rational transfer of authority or irresponsible abdication of autonomy?' *Ethics and Information Technology* 1(3): 173-184.

Lanzing, M. (2018) '"Strongly Recommended" Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies.' *Philosophy & Technology* https. doi.org/10.1007/s13347-018-0316-4.

Latonero, M (2019) *Governing Artificial Intelligence: Upholding Human Rights and Human Dignity*, Data & Society. Available at https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf (Accessed 6 May 2019).

Leonelli , S (2018) 'Rethinking Reproducibility as a Criterion for Research Quality' in Fiorito, L., Scheall, S., and Suprinyak, C.E. (eds.) *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise (Research in the History of Economic Thought and Methodology, Volume 36B)* Emerald Publishing Limited, 129 – 146.

Liu, H.Y. (2016) 'Refining responsibility: Differentiating two types of responsibility issues raised by autonomous weapons systems'. Edited by N. Bhuta, S. Beck, R. Geiss, H.Y. Liu, and C. Kress. *Autonomous weapons systems—Law, ethics policy* at 325-344. CUP: New York.

Liu, H.Y. and Zawieska, K. (2017) 'From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence.' *Ethics and Information Technology*. 19(3):1-13.

Lodge, M. and K. Wegrich (2012). *Managing Regulation*. London, Palgrave Macmillan.

Lohr, J. Maxwell, W. and Watts, P. (2019) 'Legal practitioners' approach to regulating AI risks.' In *Algorithmic Regulation*. Edited by K. Yeung & M. Lodge. OUP: Oxford. In press.

Loui, M. C. and Miller, K. W. (2008). 'Ethics and Professional Responsibility in Computing'. In *Wiley Encyclopedia of Computer Science and Engineering*. Edited by B. W. Wah. doi:10.1002/9780470050118.ecse909

Lunney, M and Oliphant, K (2013) *Tort Law*. 5th edition. Oxford University Press: Oxford.

Mangan, D. (2017) 'Lawyers could be the next profession to be replaced by computers.' 17 February. Available at at https://www.cnbc.com/2017/02/17/lawyers-could-be-replaced-by-artificial-intelligence.html (Accessed 6.11.18).

Mantelero, A. (2018) 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment.' *Computer Law & Security Review* 34(4): 754-772.

Mantalero, A. (2019), *Artificial Intelligence and Data Protection: Challenges and Possible Remedies.* Report prepared for Council of Europe, Consultative Committee of the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, T-PD(2018)09Rev. Guidelines on

Artifical Intelligence and Data Protection. Available at https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6 (Accessed 6 May 2019).

Matthias, A. (2004) 'The responsibility gap: Ascribing responsibility for the actions of learning automata.' *Ethics and information technology* 6(3): 175-183.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big Data–A Revolution That Will Transform How We Live, Think and Work*. London, John Murray.

Menn, J. and D. Volz (2016) 'Exclusive: Google, Facebook quietly move toward automatic blocking of extremist videos.' Available at https://www.reuters.com/article/us-internet-extremism-video-exclusive-idUSKCN0ZB00M. (Accessed 7.11.18)

McSherry, M. (2018) 'Will AI Widen or Weaken the Global Digital Divide?' *Medium,* 21 May (Accessed 1 May 2019).

Merton, R. K. (1942) 'The Normative Structure of Science'. In *The Sociology of Science: Theoretical and Empirical Investigations*. Edited by R. K. Merton. Chicago, Il, University of Chicago Press: 267-278.

Metcalf, J., & Crawford, K. (2016). 'Where are human subjects in Big Data research? The emerging ethics divide' *Big Data & Society*. https://doi.org/10.1177/2053951716650211

Metzinger (2019) 'Ethics Washing Made in Europe', *Der Taggespiegel*, 8 April. Available at https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html (Accessed 6 May 2019).

Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (eds.) (2013) *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Miller, A.A. (2014) 'What Do We Worry about When We Worry about Price Discrimination -The Law and Ethics of Using Personal Information for Pricing' *J. Tech. L. & Pol'y* 19:41.

Morgan, B. and Yeung, K., 2007. *An Introduction to Law and Regulation: Text and Materials*. Cambridge: Cambridge University Press.

Moses, L.B. and Koker, L.D. (2017) 'Open secrets: Balancing operational secrecy and transparency in the collection and use of data by national security and law enforcement agencies.' *Melb. UL Rev.* 41: 530.

Narula, G. (2018) 'Everyday Examples of Artificial Intelligence and Machine Learning.' 29 October. Available at https://www.techemergence.com/everyday-examples-of-ai/ (Accessed 7.11.18)

Nemitz, P. (2018). 'Constitutional Democracy and Technology in the Age of Artificial Intelligence'. *Phil Trans. A*. 376

Nevjans, N. (2016) European Parliament, Legal and Parliamentary Affairs Committee (2016). *European Civil Law Rules for Robotics.* Study for the JURI Committee. Available at http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf (Accessed 7.11.18)

Nilsson, N.J. (2014) *Principles of artificial intelligence*. Morgan Kaufmann

Nissenbaum, H. (1996) 'Accountability in a computerized society.' *Science and Engineering Ethics*. 2(1): 25-42.

Nissenbaum, H. (1996). 'Accountability in a Computerized Society.' *Science and Engineering Ethics* 2: 25-42.

Nissenbaum, H. (2010) *Privacy in Context: Technology, Policy and the Integrity of Social Life*. Stanford CA: Stanford Law Books.

Nissenbaum, H. (2011) 'A contextual approach to privacy online.' *Daedalus the Journal of the American Academy of Arts & Sciences*. 140(4): 32–48.

Noto La Diega, G. (2018) 'Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection and Freedom of Information.' J*ournal of Intellectual Property, Information Technology and E-Commerce Law*. 9 (3): 11-16..

Nuffield Foundation and the Leverhulme Centre for the Future of Intelligence (2019) *Ethical and Social Implications of Algorithms, Data and Artificial Intelligence: A Roadmap for Research*. Available at https://www.adalovelaceinstitute.org/nuffield-foundation-publishes-roadmap-for-ai-ethics-research/ (Accessed 6 May 2019).

Oberdiek, J. (2017) *Imposing risk: a normative framework*. Oxford: Oxford University Press.

Olsen, M. (1965). *The Logic of Collective Action - Public Goods and the Theory of Groups*. Cambridge, MA, Harvard University Press.

Oliver, D., (1994) 'Law, politics and public accountability. The search for a new equilibrium.' *Public Law* 238-238.

O'Neil, C. (2016) *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Oswald, M., Grace, J., Urwin, S. and Barnes, G.C., (2018) 'Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality' *Information & Communications Technology Law*. 27(2):223-250.

Oxera (2018) *Consumer Data in Online Markets*. Paper prepared for Which? 5 June.

Pasquale, F. (2015) *The Black Box Society: The secret algorithms that control money and information*. Harvard University Press.

Pariser, E. (2012). *The Filter Bubble*. London, Penguin Books.

Pasquale, F (2015). *The Black Box Society*. Boston: Harvard University Press.

Pichai, S. (2018) 'AI at Google: our principles'. 7 June. Available at: https://www.blog.google/technology/ai/ai-principles/ (Accessed 6.11.18)

Polyakov, A. (2018) 'Seven Ways Cybercriminals Can Use Machine Learning'. Available at https://www.forbes.com/sites/forbestechcouncil/2018/01/11/seven-ways-cybercriminals-can-use-machine-learning/#1e42a2a81447 (accessed 6.11.18)

Power, M. (1997). *The Audit Society*. Oxford: Oxford University Press.

Powles, J. (2015). 'We are citizens, not mere physical masses of data for harvesting'. *The Guardian*. 11 March. Available at https://www.theguardian.com/technology/2015/mar/11/we-are-citizens-not-mere-physical-masses-of-data-for-harvesting (Accessed 5.11.18)

Prainsack, B. (2019). Logged out: Ownership, exclusion and public value in the digital data and formation commons.' *Big Data & Society*. https://doi.org/10.1177/2053951719829773.

Rainey, B., Wicks, E. and Ovey, C. (2014) *Jacobs, White and Ovey: the European Convention on Human Rights* (6th edition). Oxford: Oxford University Press.

Raso, F., Hilligoss, H,, Krishnamurthy, V., Bavitz,C., & Kim,. (2018) *Artificial Intelligence & Human Rights: Opportunities & Risks.* Berkman Klein Center for Internet & Society, Harvard University. Available at https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights (Accessed 5.11.2018)

Raz, J. (1986) *The Morality of Freedom.* Oxford: Oxford University Press.

Rieder, B. (2016) 'Big data and the paradox of diversity'. *Digital Culture & Society* 2(2):39-54.

Risse, M. (2018) 'Human Rights and Artificial Intelligence: An Urgently Needed Agenda'. *Harvard Kennedy School Faculty Research Working Paper Series*. RWP18-015. Available at https://research.hks.harvard.edu/publications/getFile.aspx?Id=1664 (Accessed 6.11.18)

Royvroy, A. (2016) '"Of Data and Men." Fundamental Rights and Freedoms in a World of Big Data'. *Report for the Bureau of the Consultative Committee of the Convention for the Protection of Individuals With Regard to Automatic Processing of Personal Data*, Council of Europe, TD-PD-BUR. Available at https://rm.coe.int/16806a6020 (Accessed 6.11.18).

The Royal Academy of Engineering (2009) *Autonomous Systems: Social Legal and Ethical Issues.* August. Available at https://www.raeng.org.uk/publications/reports/autonomous-systems-report (Accessed 6.11.18).

The Royal Society (2017). *Machine Learning: The power and promise of computers that learn by example.* April. Available at https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf (Accessed 6.11.18).

Russell, S.J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia: Pearson Education Limited.

SAE International (2018) *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. Available at https://www.sae.org/standards/content/j3016_201806/ (Accessed 6.11.18).

Samek et al (2017). 'Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.' *ITU Journal: ICT Discoveries*, Special Issue No.1: 1-10.

Sandvig, C., Hamilton, K., Karahalios, K. and Langbort, C., (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 1-23. Available at http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf (Accessed 7.11.18)

Schut, M. and Wooldridge, M. (2000) June. Intention reconsideration in complex environments. In *Proceedings of the fourth international conference on Autonomous agents*. 209-216. ACM.

Schwab, K., Davies, N. and Nadella, S. (2018) *Shaping the Fourth Industrial Revolution*. World Economic Forum.

Scott, M. and Isaac, M. (2016) 'Facebook Restores Iconic Vietnam War Photo It Censored for Nudity'. Available at: https://www.nytimes.com/2016/09/10/technology/facebook-vietnam-war-photo-nudity.html (Accessed 6.11.18)

Shadbolt, N. and Hampson, R. (2018) *The Digital Ape*. Scribe: Melbourne.

Skilton, M. and Hovsepian, F. (2017). *The 4th Industrial Revolution: Responding to the Impact of Artificial Intelligence on Business*. Springer.

Smith, A. (2018) 'Franken-algorithms: the deadly consequences of unpredictable code'. *The Guardian*. 30 August. Available at https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger (Accessed 6.11.18)

Solove, D.J. (2012) 'Introduction: Privacy self-management and the consent dilemma'. *Harvard Law Review* 126: 1880.

Solum, L.B. (1991) 'Legal personhood for artificial intelligences' *NCL Rev.*70: 1231.

Sparrow, R. (2007) 'Killer Robots', *Journal of Applied Philosophy* 24(1): 62.

Su, Xiaoyuan, and Taghi M. Khoshgoftaar (2009) 'A survey of collaborative filtering techniques.' *Advances in artificial intelligence*.

Swan, M. (2015). Connected car: quantified self becomes quantified car. *Journal of Sensor and Actuator Networks* 4(1): 2-29.

Sullins, J.P. (2005) 'Ethics and artificial life: From modelling to moral agents' *Ethics and Information technology*, *7*(3), p.139.

Taplin, J. (2018) *Move fast and break things*. Politikens Forlag.

Teubner, G (2006) 'Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law.' *Journal of Law and Society.* 33: 497-521.

Tebuner, G (2018) 'Digital Personhood? The Status of Autonomous Software Agents in Private Law'. Available via SSRN network (Accessed 21.5.2019).

Thomas, M (2015) 'Should We Trust Computers?' *Gresham Lectures*: London. 20 October. Available at https://www.gresham.ac.uk/lectures-and-events/should-we-trust-computers (Accessed 3 May 2019).

Thomas, M (2017a) 'Safety Critical Systems'. *Gresham Lectures*: London. 10 January. Available at https://www.gresham.ac.uk/lectures-and-events/safety-critical-systems (Accessed 3 May 2019).

Thomas, M (2017b) 'Is Society Ready for Driverless Cars?' Gresham Lectures, London, 24 October. Available at https://www.gresham.ac.uk/lectures-and-events/is-society-ready-for-driverless-cars (accessed 3 May 2019);

Thompson, D. (1980) 'Moral Responsibility of Public Officials: The Problem of Many Hands', *The American Political Science Review* 74(4): 905-916. doi:10.2307/1954312

The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems (2018). Available at https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf (Accessed 6.11.18)

Townley, C., Morrison, E., & Yeung, K. (2017) 'Big Data and Personalized Price Discrimination in EU Competition Law'. *Yearbook of European Law* 36(1): 683-748.

Tufekci, Z. (2015) 'Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency'. *J. on Telecomm. & High Tech. L.* 13: 203.

UK Competition and Markets Authority (2018) *Pricing Algorithms*. 8 October. CMA 94. Available at

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/746353/Algorithms_econ_report.pdf (Accessed 3 May 2019).

UK Information Commissioner's Office (2018) *Democracy Disrupted – Personal Information and Political Influence*. 11 July. Available at https://ico.org.uk/media/2259369/democracy-disrupted-110718.pdf (Accessed 3 May 2019).

UK Government (2019) *Online Harms White Paper*, CP 57. Available at https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper-executive-summary--2 (Accessed 3 May 2019).

UK Department for Business, Energy and Industrial Strategy (2018) *Artificial Intelligence Sector Deal*. Available at https://www.gov.uk/government/publications/artificial-intelligence-sector-deal (Accessed 6.11.18).

UK Department for Digital, Culture, Media and Sport (2018) 'Up to £50 million to develop world leading AI talent in the UK'. Available at https://www.gov.uk/government/news/up-to-50-million-to-develop-world-leading-ai-talent-in-the-uk (Accessed 6.11.18).

UK House of Commons, Digital Culture Media and Sports Committee, *Disinformation and 'fake news' : Final Report*, Eighth Report of Session 2017-2019, 14 February, HC 1791. Available at https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf (Accessed 6 May 2019).

Universite de Montreal (2017) *Montreal Declaration for the Responsible Development of AI*. Available at https://www.montrealdeclaration-responsibleai.com/the-declaration (Accessed 6 May 2019).

UN General Assembly (2018) Report by the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. Seventy-third session. 29 August. A/73/348. Available at https://freedex.org/wp-content/blogs.dir/2015/files/2018/10/AI-and-FOE-GA.pdf (Accessed 7.11.18)

UN Special Representative of the Secretary General (2011) *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*. Endorsed by the UN Human Rights Council by Resolution 17/4 of 16 June 2011. Available at https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf. (Accessed 7.11.18.)

U.S. Citizenship and Immigration Service (2018) *Meet Emma, Our Virtual Assistant*. Available at https://www.uscis.gov/emma (Accessed 6.11.18).

U.S. Department of Transportation (2017) *Automated Driving Systems: A Vision for Safety 2.0*. Available at https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf (Accessed 6.11.18)

Vaidhyanathan, S. (2011). The Googlization of Everything: (And Why We Should Worry). University of California Press.

Van der Sloot, B. (2014) 'Do data protection rules protect the individual and should they? An assessment of the proposed General Data Protection Regulation'. *International Data Privacy Law* 4(4):307-325.

Van Est, R. and J.B.A. Gerritsen, with the assistance of L. Kool (2017) *Human rights in the robot age: Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality*. Expert report written for the Committee on Culture, Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE). The Hague: Rathenau Instituut. Available at https://www.rathenau.nl/sites/default/files/2018-02/Human%20Rights%20in%20the%20Robot%20Age-Rathenau%20Instituut-2017.pdf. (Accessed 6.11.18).

Veale, M. and Binns, R. (2017) 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data' *Big Data & Society* 4(2) doi: 10.1177/2053951717743530.

Wagner, B. (2017) 'Study On The Human Rights Dimensions of Automated Data Processing Techniques (In Particular Algorithms) And Possible Regulatory Implications'. 6 October. Council of Europe, Committee of Experts on internet intermediaries (MSI-NET). Available at https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a (Accessed 6.11.18)

Wagner, B. (2019) 'Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?' in *Being Profiling: Cogitas Ergo Sum.* Edited by M Hildebrandt. Amsterdam University Press, Amsterdam. Forthcoming.

Wallace, R.J. (1994) *Responsibility and the Moral Sentiments*. Harvard University Press: Boston.

Watson, G. (2004) *Agency and Answerability: Selected Essays*. Clarendon Press: Oxford.

Weller, A. (2017) 'Challenges for transparency'. Paper presented at 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017), Sydney, NSW, Australia. Available at *arXiv preprint arXiv:1708.01870*. (Accessed 6.11.18).

Which? (2018) *Control, Alt or Delete? Consumer research on attitudes to data collection and use.* Policy Research Report. June.

White, A. (2018) 'EU calls for $24 billion in AI to keep up with China, U.S.' Available at https://www.bloomberg.com/professional/blog/eu-calls-24-billion-ai-keep-china-u-s/ (Accessed 6.11.18).

Wierzynski, C. (2018) 'The Challenges and Opportunities of Explainable AI' 12 January. Available at https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/ (accessed 27.3.18).

Yao, M. (2017) 'Chihuahua or muffin? My search for the best computer vision API'. Available at https://medium.freecodecamp.org/chihuahua-or-muffin-my-search-for-the-best-computer-vision-api-cbda4d6b425d (Accessed 6.11.18).

Yearsley, L. (2017). "We Need to Talk About the Power of AI to Manipulate Humans." 5 June. *MIT Technology Review.* 5 June. Available at https://www.technologyreview.com/s/608036/we-need-to-talk-about-the-power-of-ai-to-manipulate-humans/ (Accessed 7.11.18).

Yeung, K., (2011) 'Can we employ design-based regulation while avoiding brave new world?' *Law, Innovation and Technology*. 3(1): 1-29.

Yeung, K. (2015) 'Design for Regulation.' In *Handbook of Ethics, Values and Technological Design,* edited by M. J. Van Den Hoven, P.E. Varmaas and I. van de Poel. Dordecht: Springer.

Yeung, K. (2016) '"Hypernudge": Big Data as a mode of regulation by design'. *Information, Communication & Society:* 1-19.

Yeung, K. (2017a) 'Algorithmic regulation: a critical interrogation', *Regulation & Governance.* doi 10.1111/rego.12158.

Yeung, K., (2017b) 'Blockchain, Transactional Security and the Promise of Automated Law Enforcement: The Withering of Freedom Under Law?' In *3THICS - The Reinvention of Ethics in a Digital Age,* Edited by P. Otto and E. Graf at 132-146.

Yeung, K. (2018a) 'Five Fears About Mass Predictive Personalization in an Age of Surveillance Capitalism'. *International Data Privacy Law* 8: 258-269

Yeung, K. and Weller, A. (2018b). 'How is 'transparency' understood by legal scholars and the machine learning community?' *Being Profiling. Cogitas Ergo Sum*. In E. Bayamlioglu, I. Baraliuc,

L. W. Janssens and M. Hildebrandt (eds.) Amsterdam University Press..

Zalnieriute, M., et al. (2019). "The Rule of Law and Automation of Government Decision-Making." Modern Law Review 82: 397-424.

Zliobaite, I. (2015) A survey on measuring indirect discrimination in machine learning. Available at *arXiv preprint arXiv:1511.00148* (Accessed 6.11.18).

Zook, M. and Grote M. H. (2017). 'The microgeographies of global finance: High-frequency trading and the construction of information inequality' *Environment and Planning A: Economy and Space* 49(1): 121-140

Zuboff, S., (2015) 'Big other: surveillance capitalism and the prospects of an information civilization' *Journal of Information Technology* 30(1): 75-89.

Zweig, K. A., Wenzelburger, G. and Krafft, T. D (2018) 'On Chances and Risks of Security Related Algorithmic Decision-Making Systems'. *European Journal for Security Research*. 3(2):181-203.

******

Advanced digital technologies and services, including AI tools, come with extraordinary promise, particularly in the form of enhanced efficiency, accuracy, timeliness and convenience across a wide range of services. Yet the emergence of these technologies is also accompanied by rising public anxiety regarding their potentially damaging effects for individuals, for vulnerable groups and for society more generally.

Given their pervasiveness in daily life, we must acquire a deeper understanding of their impact on the exercise of human rights and fundamental freedoms, and we should carefully consider how to allocate responsibility in case of adverse consequences. If we are to take human rights seriously in a globally connected digital age, we cannot allow the power of our advanced digital technologies and systems, and those who wield and derive benefits from them, to be accrued and exercised without responsibility.

Effective and democratically legitimised governance arrangements and enforcement mechanisms must be put in place to ensure that responsibility for the risks, harms and wrongs arising from the operation of advanced digital technologies are duly allocated.

www.coe.int/freedomofexpression

PREMS 111819

ENG

www.coe.int

The Council of Europe is the continent's leading human rights organisation. It comprises 47 member states, including all members of the European Union. All Council of Europe member states have signed up to the European Convention on Human Rights, a treaty designed to protect human rights, democracy and the rule of law. The European Court of Human Rights oversees the implementation of the Convention in the member states.

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE