Assembling the Algorithm for Human Rights Centred AI Regulation

Chloe Haden

Submitted to the University of Hertfordshire in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

University Of Hertfordshire
School of Law

February 2024

Abstract:

This thesis argues that inspiration should be taken from the General Data Protection Regulation to form the basis of future Artificial Intelligence (AI) regulation in England and Wales. Currently, AI poses significant challenges to society, undermining human rights, conflicting with liability frameworks, and many systems having no lawful basis of use. Through the proposed introduction of AI data protection principles, pre-market and post-market monitoring requirements, and rules to clarify enforceability and accountability mechanisms, this thesis gives recommendations to achieve a human rights focused AI framework, clarifying issues of liability and strengthening human right protections. By utilising a thematic analysis of extensive literature, and a doctrinal methodology to examine case-law, regulatory proposals and legislation, a techno-legal-ethics interdisciplinary perspective is taken to provide a unique insight into AI regulation. Through examination of the most recent regulatory developments in the European Union (EU), the suggestions proposed seek to take advantage of the mistakes made, to correct areas where future EU regulation falls short, and to offer solutions to provide stronger safeguards to human rights.

By providing an alternative perspective on how to regulate AI, this research seeks to contribute to the field of AI and legal research, and provide an invaluable resource for those examining the key areas of concern and recent developments in regulating AI. These proposals seek to aid policymakers in considering alternatives to ideas already presented in the field, to ensure priority is given to the safeguarding of human rights. The suggestions intend to form the basis of blanket regulation, allowing the recommendations to be built upon in the future, whilst also opening discussion to further avenues of research, and further legislation in the future. For society to reap the benefits of AI, effective regulation that prioritises human rights must be introduced.

Acknowledgements:

I would like to express my sincere gratitude to my supervisory team who have helped tremendously in guiding me during the years of completing this thesis:
Thank you to Dr. Felipe Moreno-Romero who has provided me with support and feedback through every stage of the research project, your advice has been invaluable, and I'm very thankful for the opportunities you have given me to be involved in your own research.
I would also like to thank Professor Farshid Amirabdollahian for his insightful comments and contributions to help me develop my understanding of the technical side of the research, your advice and support has been greatly appreciated, and helped to promote technical accuracy within my project. Also, I would like to express thanks to Professor Constantine Sandis, who helped to shape my research in the preliminary stages, and who also provided me with great support and advice, ensuring that my arguments were well balanced from all angles.

I would also like to thank the University of Hertfordshire for granting me a studentship to support me financially for the preliminary years of this research.

On a personal note, I would like to thank my Mum, for always being my number one supporter, and giving unwavering confidence in my ability to achieve my goals.

I would also like to thank several others, including my family, friends, colleagues, and students, who have supported, encouraged, and cheered me past the finish line!

Table of Abbreviations

AGI	Artificial General Intelligence
AI	Artificial Intelligence
ANI	Artificial Narrow Intelligence
ANNs	Artificial Neural Networks
CoE	Council of Europe
DPA	Data Protection Act
DPD	Data Protection Directive
DPIA	Data Protection Impact Assessment
EC	European Commission
ECHR	European Convention on Human Rights
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EP	European Parliament
EU	European Union
EUAIB	European Union Artificial Intelligence Board
FRIA	Fundamental Rights Impact Assessment
FRT	Facial Recognition Technology
GDPR	General Data Protection Regulation
HLEG	High-Level Expert Group
ICO	Information Commissioner’s Office
IP	Intellectual Property
ML	Machine Learning
MS	Member States
NHS	National Health Service
NLP	Natural Language Processing
PLD	Product Liability Directive
RQ1	Research Question 1
RQ2	Research Question 2
RQ3	Research Question 3
UDHR	Universal Declaration of Human Rights
UN	United Nations

Glossary of Terms

AI Accuracy: Measured by the total number of correct predictions made within the decision-making process, in reference to the number of false positives and false negatives.

Article 22: A data subject right included within the General Data Protection Regulation (GDPR) not to be subject to a decision based solely on automated decision-making of profiling.

Artificial General Intelligence (AGI): AI systems that pass the Turing Test. Currently, it is argued that AI is yet to reach this level.

Artificial Intelligence (AI): Products or components that have differing degrees of intelligence and autonomy, which receive data inputs to produce outputs, and can perform one or more specific human tasks.

Artificial Narrow Intelligence (ANI): The level of intelligence AI has reached. An algorithm that is designed to be an expert at one, or few tasks.

Artificial Neural Networks (ANNs): Commonly used for solving business problems and are trained on the basis and idea of the brain, to model complex patterns and predictions.

Artificial Super Intelligence (ASI): The level of intelligence that surpasses human intelligence. It is heavily debated whether this much intelligence will ever exist.

Black Box: AI systems which lack transparency, where AI developers are often unable to understand or explain the algorithm outputs.

Deployers: The authorities, bodies or companies who use AI technology and systems.

Developers: The manufacturers and creators of AI technology and systems.

Expert Systems: Programs that contain expert knowledge in a specific area to assist professionals.

Explainability: In this context, where there is the ability to trace back how and why a particular AI decision was reached.

Fuzzy Logic: Programs that can process all available options to make quick decisions, regardless of the data set size, and is commonly used in combination with ANNs and NLP

Interpretability: Making transparency understandable to those deploying, using or being subject to AI decisions.

Machine Learning (ML): the idea that systems can learn from data, solve problems, and make decisions with minimal human interaction.

Mandatory sources: Legal sources which are binding to the domestic jurisdiction, in this context, England and Wales.

Natural Language Processing (NLP): A machine learning technique that allows understanding and interpretation of human language.

Persuasive Sources: Non-binding legal sources, or secondary sources that can persuade and influence opinion. The Courts can follow these sources, but they are not obliged to do so.

Profiling: Predictions or decisions made from an analysis of an individual's characteristics and traits.

Right to an explanation: Where explainability is mandated by law.

Transparency: In this context, the ability to see and understand the decision-making processes that AI systems complete.

Turing Test: A test established by Alan Turing, to assess the intelligence of machines. To pass the Turing Test, systems would need to be completely indistinguishable from human counterparts.

Table of Contents

Chapter 1: Introduction

This thesis answers the overarching question: 'to what extent should Artificial Intelligence (AI) be regulated to ensure an ethical framework centred on human rights?'. When considering this question, one must look to other recent and effective legislation that has been impactful and influential. For this reason, this thesis argues that using the General Data Protection Regulation (GDPR)[1] as a basis for future AI regulation would be the best approach to ensure a human rights centred framework.

The ultimate research aim is to produce solutions to AI regulation which are up to date, well-researched, justified, and realistic, conforming to already enacted legislation applicable in England and Wales, including the Data Protection Act (DPA),[2] and the European Convention on Human Rights (ECHR).[3] The DPA is an implementation of the European Union's (EU) GDPR, which although subject to criticism,[4] was brought into force in 2018, and covers the broad topic of data protection and safeguarding personal data. The ECHR was brought into force in 1953 by the Council of Europe (CoE),[5] drawing inspiration from the Universal Declaration of Human Rights (UDHR),[6] and introduced as a focused effort to uphold human right protections throughout the related Member States (MS).

To achieve the overarching aim noted above, the thesis has the first objective to explore the issues AI systems pose to ethics, human rights, and current liability frameworks, to suggest ways in which a new framework could address these problems. The second objective requires an assessment of the impact made by the

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

[2] Data Protection Act 2018.

[3] Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14 1950 (ECHR).

[4] Frederik Borgesius, 'Strengthening legal protection against discrimination by algorithms and artificial intelligence' [2020] 24(10) *I.J.H.R Rights* 1; Cambridge Consultants, *Use of AI in Online Content* Moderation (on behalf of Ofcom, 2019) 55; Fabienne Ufert, 'AI Regulation Through the Lens of Fundamental Rights: How Well Does the GDPR Address the Challenges Posed by AI?' [2020] 5(2) *European Papers* 1087; Sam Wrigley, 'Taming Artificial Intelligence: "Bots," the GDPR and Regulatory Approaches' in Corrales M. Fenwick M. and Forgó N. (eds) *Robotics, AI and the Future of Law. Perspectives in Law, Business and Innovation* (Springer Singapore, 2018) 185-186.

[5] ECHR (n 3).

[6] Universal Declaration of Human Rights (adopted 10 December 1948 UNGA Res 217 A(III) (UDHR).

GDPR on AI systems, and whether there is a need to reform the DPA[7] to achieve alignment with the future of AI. The final objective requires analysis of the key regulatory issues concerning AI, to recommend solutions to achieve a new human rights centred framework for AI. This research aspires to contribute to solving the issues regarding the regulation of AI, which whilst relatively new, is an increasingly trending issue. This tackles a current research area which affects the law, ethics, technology, and society. The thesis draws these perspectives together to not only add to existing knowledge, but also build to the literature directly focused on the regulation of AI. This research will demonstrate that reform of the DPA, in combination with a new framework for AI will have a significant positive effect on AI ethical and technological progression. The solutions produced comply with the standards within the DPA and the ECHR, and have an ethical and realistic focus, including requirements, monitoring, and obligations in law for both developers and deployers of AI systems.

1.1 Background

1.1.1 What is AI?

The term ‘Artificial Intelligence’, originally coined in 1955 by cognitive scientist and computer science pioneer John McCarthy,[8] has since been used as an umbrella term for various high intelligence machines, with capabilities such as machine learning (ML), computer vision, visual perception and decision-making.[9] Following the end of World War Two, scientists from various disciplines were brought together to tackle the challenges of intelligent machines in Britain.[10] McCarthy supported this effort, and defined the term AI as *“making a machine behave in ways that would be called intelligent if a human being were behaving*”.[11] McCarthy understood the term as the science and engineering of making intelligent machines, especially intelligent computer programs.[12]

[7] Data Protection Act 2018.

[8] J. McCarthy, M.L. Minsky, N. Rochester and C.E. Shannon, ‘A Proposal For the Dartmouth Summer Research Project on Artificial Intelligence’ (1955) <https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html> accessed 12th February 2020.

[9] Mark Howard, *Artificial Intelligence, Machine Learning and Deep Learning,* (CreateSpace Publishing, 2018) 200.

[10] Ronald Chrisley, *Artificial Intelligence: Critical Concepts, Volume 1* (Taylor & Francis, 2000) 343.

[11] McCarthy, Minsky, Rochester and Shannon (n 8).

[12] ibid.

There have been various definitions of AI since,[13] including "*machines that respond to stimulation consistent with traditional responses from humans, including the capacity for contemplation, judgement and intention*",[14] and put more simply, "*the field of computer science dedicated to solving cognitive problems commonly associated with human intelligence*".[15] Defining computer intelligence has been a subject of debate for decades, dating back to the 1950s, before the term AI had even been introduced. The Turing Test was established in 1950, named after Alan Turing, a famous mathematician and computer scientist, and gave a practical solution for defining computer intelligence, by seeking whether a human evaluator would be able to distinguish between a human and a computer participant; if the evaluator failed to note the difference between the participants, the computer was seen to pass the test of computer intelligence.[16] Turing himself was said to be influenced by Ada Lovelace, considered to be the first computer programmer, and lived during the 19th Century. She is recognised due to her recognition of computers having the ability to follow a series of instructions to perform a complex calculation.[17]

Looking to more recent definitions, Ofcom refers to "*the capability of a machine to exhibit human-like performance at a defined task*".[18] The High-Level Expert Group (HLEG) on AI, set up by the European Commission (EC) has defined AI as "*systems that display intelligent behaviour, by analysing their environment and reacting, with some degree of autonomy, to achieve specific goals*".[19] The HLEG has suggested several definitions on behalf of the EC, each bringing more understanding of the term. One of the more detailed defines AI as "*software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition,*

[13] T.J.M. Bench-Capon and Paul Dunne, 'Argumentation in artificial intelligence' [2007] 171(10*) Artificial Intelligence* 619 ; Jerry Kaplan, *Artificial Intelligence: What Everyone Needs to Know* (Oxford University Press, 2016) 1-2.

[14] Shukla Shubhendu and Jaiswal Vijay, 'Applicability of Artificial Intelligence in Different Fields of Life' [2013] 1(1) *International Journal of Scientific Engineering and Research* 2347.

[15] Amazon, 'What is Artificial Intelligence? Machine Learning and Deep Learning' (Amazon website) <www.aws.amazon.com/machine-learning/what-is-ai/> accessed 12th July 2019.

[16] Alan Turing, 'Computing Machinery and Intelligence' [1950] 59(236) *Mind* 433.

[17] Octavia Reeve, 'Celebrating Ada Lovelace Day: what Ada means to us' (Ada Lovelace Institute, 8th October 2019) <https://www.adalovelaceinstitute.org/blog/celebrating-ada-lovelace-day/> accessed 15th November 2020.

[18] Cambridge Consultants (n 4) 4.

[19] High-Level Expert Group on Artificial Intelligence, *A Definition of AI: Main Capabilities and Disciplines* (Independent group set up by the European Commission, 2019) 1.

interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal".[20] This is arguably one of the more extensive definitions, and highlights the continued difficulty in getting the definition of AI just right.

The HLEG categorises certain aspects of AI into a scientific discipline, such as ML and machine reasoning, and a technology discipline, such as robotics.[21] This clarification intended to prevent misunderstanding by academics and laymen alike, and was hoped to be applied consistently within ethics guidelines and policy recommendations.[22] In the HLEG on AI's report, written on behalf of the EC, they state that AI can be purely software-based, such as voice assistance and facial recognition, or embedded as components within hardware devices, such as advanced robotics and autonomous vehicles.[23]

The majority of definitions for AI are aligned around the concept of computer programs or machines, which are capable of exhibiting behaviour regarded as 'intelligent'.[24] The most intelligent AI machines have been defined in the engineering field as consisting of two elements, the first relating to autonomy and the level of human interaction needed for the machine to operate.[25] This can be measured on a spectrum, in which the capacity of decision-making by the machine correlates with the reduction in the need for human intervention and interaction,[26] and is considered by many as the most important ability AI has.[27] The second element is intelligence, which can consist of adapting behaviour to fit new circumstances, and possessing the ability to learn, reason, and understand language.[28] These factors of autonomy and adaptiveness are key elements that differentiate AI from standard computer software. Unlike traditional software, AI is regularly involved in decision-making

[20] ibid 6.
[21] ibid 1.
[22] ibid 1.
[23] ibid 1.
[24] Kaplan (n 13) 1.
[25] Peter Asaro, 'A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics' in Lin, P. Abney, K. and Bekey, G. (eds) *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press, 2012) 169 ; Royal Academy of Engineering, *Autonomous Systems: Social, Legal and Ethical Issues* (2009) 4-9.
[26] Royal Academy of Engineering (n 25) 2.
[27] Samir Chopra and Laurence White, *A Legal Theory for Autonomous Artificial Agents* (University of Michigan Press, 2011) 10.
[28] Alex Jaimes, 'Computer vision startups tackle AI' [2016] 23(4) *IEEE* 94.

processes and can possess ML capabilities, allowing systems to learn, adapt and decide based on what is inputted into the system. Due to these unique factors, future regulation is needed to regulate the outputs of such systems, which have raised concerns in areas including bias, transparency and explainability.

The HLEG categorises AI as narrow, where AI systems can perform one or few specific tasks, or general, which includes systems that can perform the majority of activities humans do.[29] Currently, deployed systems are clarified to be narrow, and that there are *"still many open ethical, scientific and technological challenges to build the capabilities needed to achieve general AI, which would include common sense reasoning, self-awareness and the ability to define its own purpose*".[30] Even so, AI has continued to make rapid progress, and has developed worldwide, ranging in areas from computer vision to gameplay, assisting in fields of law, business and medicine, in addition to many other sectors.[31] These developments have allowed the concept of AI to be more readily accepted and discussed by academics and the technology industry, leading to a rapid and significant progression worldwide.

Understanding the evolution regarding the definition of AI is important in framing the legal, ethical and regulatory challenges examined within this thesis. As AI systems have developed, ambiguous and broad definitions have complicated efforts to produce regulation, as seen in the discussions through developing the AI Act. The ambiguity in defining AI emphasises the issue of legal uncertainty, particularly in circumstances of liability, where AI should be distinguished from traditional software, due to its unique capabilities. Given the rapid growth of the understanding of AI, the scope of the definition and its capabilities, concerns grow for alignment to current frameworks, such as data protection, justifying the need to reassess current protections.

The use of the term as a 'catch-all' phrase, encompassing a wide range of technologies, ranging from ML and NLP to robotics and generative models, can raise questions about whether the term of AI is meaningful, or whether it oversimplifies the

[29] High-Level Expert Group on Artificial Intelligence (n 19) 5.
[30] ibid 5.
[31] Mike Loukides and Ben Lorica, *What is Artificial Intelligence?* (O'Reilly Publishing, 2016) 1-2.

complex and varied nature of the technology. AI can refer to systems that range for simple rule-based programs to highly sophisticated ANNs, making the term imprecise in capturing the specific functions and capabilities of the differing types of systems. Despite this, the term AI continues to be used widely in legal, ethical and regulatory discussions as a convenient shorthand. In the context of this thesis, AI is used to capture the variety of systems that fall within the term, as the regulatory recommendations are designed to address the underlying risks and challenges, regardless of the specific type of application of the technology in question.

1.1.2 Development and Deployment of AI

AI has developed progressively, leading to six main branches of AI being established, namely ML, expert systems, artificial neural networks (ANNs), fuzzy logic, natural language processing (NLP) and robotics.[32] The modern history of machines began with the creation of stored-program electronic computers, and has developed through the increased use of ML.[33] ML is based on the idea that systems can learn from data, solve problems and make decisions with minimal human interaction.[34] One of the first major uses of AI in its simplest form dates back to 1956, whereby logic theorists Newell and Simon invented a computer program called the 'thinking machine' that could solve complex problems.[35] Since then, this idea has developed and led attention surrounding the concept to grow, resulting in rapid progression worldwide.[36] Currently in healthcare for example, AI is used in the form of complex ML algorithms to approximate patient conclusions, and in robotics to aid in surgical procedures.[37]

AI in the form of expert systems are programs that contain expert knowledge in a specific area to assist professionals, and have been commonly used to make

[32] Prachi Mate, 'Branches of Artificial Intelligence' (My Road to Artificial Intelligence, Medium Blog, 25th May 2020) <https://medium.com/myroadtoartificialintelligence/branches-of-artificial-intelligence-812b8e292cdb> accessed 19th August 2020.

[33] Alice Rawsthorn, 'Genius and Tragedy at Dawn of Computer Age' *New York Times* (New York, 25 March 2012).

[34] SAS, 'Machine Learning, what it is and why it matters' (SAS Analytics and Data Science Insights) <www.sas.com/en_gb/insights/analytics/machine-learning.html> accessed 15th July 2019.

[35] Leo Gugerty, 'Newell and Simon's Logic Theorist: Historical Background and Impact on Cognitive Modelling' [2006] 50(9) *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 880.

[36] Tom Taulii, *Artificial Intelligence Basics: A Non-Technical Introduction* (Apress Publishing, 2019) 1.

[37] Erwin Loh, 'Medicine and the rise of robotics: a qualitative review of recent advances of artificial intelligence in health [2018] 2(2) *BMJ Leader* 59.

advancements in the medical, chemical and geological fields.[38] Expert systems are also already deployed in the most fundamental sectors of society, from medical to legal institutions.[39] AI has contributed to robotics by enhancing mechanical effectors, sensors and computers, leading to use and development in the areas of social care and medicine.[40]

ANNs are commonly used for solving problems and are trained on the basis and idea of the human brain, to model complex patterns and predictions.[41] ANNs and ML are often used together for image processing and character recognition, having the ability to develop technologies such as facial recognition. Fuzzy Logic can process all available options to make quick decisions, regardless of the data set size,[42] and is commonly used in combination with ANNs and NLP. The positive progression of NLP has led to the successful sale and use of products such as Apple's Siri and Amazon's Alexa.[43]

Ofcom chooses not to consider these categories when defining AI, however, they acknowledge the breakthroughs in ML and the development of ANNs.[44] They also comment on the rapid development of AI techniques over the last decade which now can routinely collect and analyse data,[45] and although this brings efficiency through outpacing human counterparts, it brings its own set of challenges. The rapid development and innovation of AI in recent years by companies such as Amazon, Google, and the other 'tech giants' has far outrun regulation in the area, resulting in exposing major challenges, such as biased AI decision-making, misinformation, and interference with human rights.

[38] Janet Vaux, 'From expert systems to knowledge-based companies: How the AI industry negotiated a market for knowledge' [2001] 15(3) *Social Epistemology* 231.

[39] Omar Ali, Wiem Abdelbaki, Anup Shrestha, Ersin Elbasi, Mohammad Alryalat and Yogesh Dwivedi, 'A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities' [2023] 8(1) *Journal of Innovation & Knowledge* 100333 ; Philip Leith, 'The rise and fall of the legal expert system' [2016] 30(3) *I.R.L.C.T* 94.

[40] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian and Kerstin Dautenhahn, 'Towards safe and trustworthy social robots: ethical challenges and practical issues' in Tapus, A., André, E., Martin, JC., Ferland, F., Ammi, M. (eds) *Social Robotics*. (ICSR, Lecture Notes in Computer Science, Springer, Cham, Vol. 9388, 2015).

[41] Roza Dastres and Mohsen Soori, 'Artificial Neural Network Systems' [2021] 21(2) *I.J.I.R* 13.

[42] Konstantina Chrysafiadi, 'The Role of Fuzzy Logic in Artificial Intelligence and Smart Applications' in Tsihrintzis, G.A., Virvou, M. and Jain, L.C. (eds) *Learning and Analytics in Intelligent Systems* (Springer, Cham, Volume 34, 2023) 26.

[43] Jaimes (n 28).

[44] Cambridge Consultants (n 4) 4.

[45] ibid 12.

1.1.3 Issue of Liability and Lack of Regulation

It is inevitable given the rapid progression of AI, that a continuance of cases will emerge, linking to the strong need for action regarding AI regulation to assist the courts. AI is already at a progressive level in the ways it challenges fundamental ethics and cultural ideas on an everyday basis, through the integration of skilful technology with humans, and the impact it has on ethics and the legal system.[46] Already in the courts, the lack of regulation was noted in the High Court judgment of *R(Bridges),*[47] which related to the use of facial recognition technology (FRT). FRT had existed long before the term AI had been coined, and has increased in use due to developments in image recognition. The judgment highlighted that there existed no lawful basis for the use of the technology, and therefore the judgment had to rely on common law principles.[48] On appeal, it was held that the use of such technology was unlawful, due to no legal framework existing, and therefore unable to comply with the 'prescribed by law' requirement in relation to the infringement of Article 8 of the ECHR.[49] This case not only highlights the need for AI regulation, but also reflects the inconsistencies in the court's response to AI, providing further justification of the need for regulation to help clarify how and when AI can lawfully be used.

The issues that stem from the 'pacing problem' between technology and regulation need to be addressed. There needs to exist a set of rules to follow, but currently, no legislation sufficiently addresses the specific issues concerning AI systems. Rules that regulate the use of AI need to be implemented, to ensure AI continues to progress and develop in the 'most ethical' way, as well as to ensure the protection of fundamental rights and freedoms, including the protection of personal data.[50] With the lack of hard legislation, several industries, including the National Health Service (NHS) have self-regulated, or have existing soft law approaches in place, which were mainly introduced to address the ethical concerns of AI systems. The NHS has taken a code of conduct approach to AI, which consists of principles outlining how the NHS should work with technology companies developing AI systems and algorithms for

[46] Braden Allenby, 'Robotics: Morals and Machines' [2012] 481 *Nature London* 26.
[47] *R(Bridges) v Chief Constable of South Wales* [2019] EWCH 2341 (Admin), [2020] 1 WLR 672.
[48] ibid [78].
[49] *R(Bridges) v Chief Constable of South Wales* [2020] EWCA Civ 1058, [2020] 1 WLR 5037 [58].
[50] Charter of Fundamental Rights of the European Union [2000] OJ C364/1 Article 8.

use in healthcare.[51] Vehicle manufacturers including Google, Volvo and Mercedes-Benz have also self-regulated, by pledging to take accountability if one of their autonomous vehicles is found at fault, intending to promote public trust, and ease the argument of liability when using their vehicles.[52] The use of soft law for AI could be advantageous due to it being easy to adopt and modify, as well as flexible in implementation.[53]

However, soft law would not be enough, and at some point, a basic and minimum regulatory binding framework must be introduced to set a foundation, upon which other non-binding principles and codes of conduct can be established. Hadfield believes that regulatory oversight is necessary, and that discussions surrounding the soft law approach diminishes the complexity of the issue.[54] It is argued that the solution to regulating AI is open and structured discussion, followed by strong, clear regulation enacted, rather than patchworks of self-regulation.[55] Castelvecci believes that self-regulation would not work for AI, due to those companies following ethical guidelines having a disadvantage in terms of innovation to those who do not, highlighting the need for universal and enforceable regulation.[56] There is also a common perception that self-regulation allows rules to be made less in the public interest and more to protect corporate interests, which could lead to a blurring of rules, and ineffective for ensuring AI progresses in an ethical manner.[57] Also, without explicit or implicit backing from Parliament, self-regulation may have the negative impact of regulatory uncertainty, which may cause businesses to delay investment decisions leading to a stifling of innovation.[58]

[51] Department of Health and Social Care, *A guide to good practice for digital and data-driven health technologies* (Government Department, 2021).

[52] Tina Bellon, ‘Liability and legal questions follow Uber autonomous car fatal accident’ (Insurance Journal, 20 March 2018) <https://www.insurancejournal.com/news/national/2018/03/20/483981.htm> accessed 24th March 2019.

[53] Elis Tarelli, ‘The Strengths and Weaknesses of Soft Law as a Source of International Financial Regulation’ (SSRN, 2009) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1467842> accessed 19th February 2020.

[54] Gillian Hadfield, ‘Rules for Robots: The Path to Effective AI Regulation’ (MIT Digital Blog, 12th June 2019) <http://ide.mit.edu/news-blog/blog/rules-robots-path-effective-ai-regulation> accessed 21st February 2020.

[55] Davide Castelvecci, ‘AI Pioneer: “The dangers of abuse are very real” (Nature Research, 4th April 2019) <https://www.nature.com/articles/d41586-019-00505-2> accessed 6th February 2020.

[56] ibid.

[57] Daniel Castro, *Benefits and Limitations of Industry Self-Regulation for Online Behavioral Advertising* (The Information Technology and Innovation Foundation, 2011) 9.

[58] ibid 9.

1.1.4 Challenges of Regulation

A factor that needs to be considered when regulating AI is whether there exists a duty of care, to follow the current law under negligence claims. Usually, the three elements that need to be proven for negligence are that: the defendant owes a duty of care, the defendant breached that duty, and the breach caused an injury or damage, and in most cases, that it is foreseeable.[59] In England and Wales, the notion of a duty of care was established in *Donoghue v Stevenson,* in which it was clarified that individuals "*must take reasonable care to avoid acts and omissions which would likely injure their neighbour*".[60] A neighbour refers to individuals who are closely and directly affected by the original act, and ought to be reasonably considered as being affected.[61]

The 'neighbour principle' is extremely broad, resulting in attempts to narrow the scope of liability in subsequent case-law.[62] In 2018, *Robinson* clarified the applicability of establishing a duty of care. The judgment made clear that when assessing whether a duty of care exists, established principles of the law of negligence should be considered, and an extension of this should only be considered in novel situations.[63] Applying the decision of *Robinson,* it is clear that the duty of care principle could present a remedy for handling scenarios involving AI. It is not uncommon for a duty of care to be considered to exist for suppliers of goods, giving light to the Product Liability Directive (PLD), whereby liability is allocated depending on whether the injured party can prove a defect in a product, that the defect caused the related damage, and that there exists a link between them. [64]

Such principles of a duty of care could therefore be an option in handling claims involving AI, but the extent and scope of such a duty needs to be addressed and possibly reconsidered. In particular, for purposes including whether AI software is to be considered a 'product' or a 'service', and the issue of foreseeability in AI systems.

[59] Maruerite Gerstner, 'Comment, Liability Issues with Artificial Intelligence Software' [1993] 33 *Santa Clara Law Review* 239; *Donoghue v Stevenson* [1932] AC 562 [619]; *Caparo Industries plc v Dickman* [1990] 2 AC 605 [632]; *The Wagon Mound no 1* [1961] AC 388.

[60] *Donoghue v Stevenson* (n 59) [580].

[61] ibid [580].

[62] *Anns v Merton London Borough Council* [1978] AC 728 [3] ; *Caparo Industries plc v Dickman* [609].

[63] *Robinson v Chief Constable of West Yorkshire Police* [2018] UKSC 4 [2018] WLR 595 [30]

[64] Council Directive (EC) 85/374 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products (Product Liability Directive) [1985] OJ L 210, 7.8.

In the EU's report on liability for AI, it is argued that compliance with an adapted range of duties of care should be mandated for operators, including choosing the appropriate system for the appropriate level of tasks and skill, and monitoring and maintaining systems.[65] However, it is noted that the more advanced technologies become, the more difficult the compliance with duties of care will be, and training will need to be given to ensure operators have the necessary expertise to understand their systems.[66] The EC recognise that although it is possible to apply existing liability regimes to emerging technologies, due to the challenges and limitations of these regimes that were formulated on older concepts, doing so may leave victims partially, or entirely uncompensated,[67] and are therefore questionable to rely on in the long-term.

Due to this, the issue between AI and foreseeability needs to be addressed. The ability to act autonomously stands out as the key feature to separate AI from other technologies, with systems already having the ability to perform tasks that require substantial skill, such as driving a car or building an investment portfolio, without direct human control or supervision.[68] If a negligence case arises based on the actions of AI, which is inevitable, courts will be faced with the difficult question of foreseeability.[69] Several AI systems are designed not only to respond to pre-defined situations, but to additionally identify and classify new ones and connect them to a self-chosen corresponding action,[70] which would be a common necessity when controlling a car. The more AI systems advance their capability of processing, the more difficult it is to foresee the precise impact they will have in operation.[71] This process is a form of ML and gives software the capability to develop and progress itself, evolving over time from its own experiences.[72]

[65] European Commission, *Liability for Artificial Intelligence and other emerging digital technologies* (Expert Group on Liability and New Technologies- New Technologies Formation, 2019) 7.
[66] ibid 45.
[67] ibid 19.
[68] Matthew Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' [2016] 29(2) *Harvard Journal of Law and Technology* 354.
[69] Weston Kowert, 'The Foreseeability of Human-Artificial Intelligence Interactions' [2017] 96(1) *Texas Law Review* 181.
[70] European Commission (n 67) 33.
[71] ibid 33.
[72] Kowert (n 71).

However, this means that when AI is placed into public use, developers will be highly unlikely to accurately predict how AI systems will solve the tasks and problems they encounter.[73] Due to this unpredictability, developers may have a level of protection in claims, especially where an AI system has acted unexpectedly after being sold.[74] This arguably places a bigger burden on consumers (whether that be the deployer or end-user), as if an AI system were to act unexpectedly and developers were somewhat protected, other influences on the AI system will receive the shift of blame, which may be the deployer or the end-user, as they would have had the greatest influence to the system after being sold. The extent and level of skill required to complete tasks given to AI will undoubtedly continue to increase in the coming years, which will force disruptive changes to the law as the legal system faces the challenge of dealing with the increasing number of autonomous machines.[75]

1.1.5 AI that Processes Personal Data and the Impact of the GDPR

As seen, AI is a large and complex area, being used within different systems, machines, programs, and algorithms, which are used widely across a variety of industries. For the purposes of this research, and to make the scope of the project manageable, the research focuses on the regulation of AI that is trained using data sets that include personal data, and/or that process personal data when in use, following the GDPR's definition of personal data included within Article 4(1).[76] This is still a substantial area, as the majority of systems require large amounts of data to learn and train from, and usually, personal data is amongst these data sets.[77] Concern has been raised in this area in regard to the issue of bias, and as discussed in the below section (1.1.8), the lack of transparency and explainability requirements continue, which highlights the assurance needed that AI systems directly comply with data protection principles, as well as the fundamental rights included within the ECHR.[78]

[73] ibid.
[74] ibid.
[75] Scherer (n 70).
[76] GDPR (n 1) Article 4(1).
[77] Anna Oleksiuk, 'How to train AI with GDPR limitations' (Intellias Intelligent Software Engineering Blog, 13th September 2019) <https://www.intellias.com/how-to-train-an-ai-with-gdpr-limitations/> accessed 20th August 2020.
[78] ECHR (n 3).

Google's CEO has reportedly stated that the GDPR can serve as a strong foundation for AI regulation.[79] This is agreed with by the United Nations (UN), who comment that sectorial legislation for AI would be preferable,[80] and recommend that existing regulation needs to be updated to ensure human rights are safeguarded, especially in the area of data protection.[81] Spyridaki expresses that due to the GDPR having the strongest impact on any law relating to the creation of a better regulated data market; with data being the basis for AI applications, regulation of AI and the GDPR should be considered together.[82] It should be recognised that the GDPR and AI regulation should not be exclusive to one another, and instead, should be complementary, to allow focus on the underlying principles relating to the GDPR, namely the protection of privacy and ethical practices.[83]

It is worth noting that future regulation designed to govern AI should take into consideration the GDPR, for reasons such as its practical implementation by organisations, and to avoid unnecessary duplicated or potentially conflicting obligations, ambiguity or legal uncertainty.[84] The GDPR affects the use of AI in at least three ways, which include limitations on the collection and use of data, restrictions on automated decision-making and an increase in compliance costs and risks.[85] The GDPR includes some provisions which already serve as a degree of regulation to AI, in particular; Article 4(4) on the definition of profiling,[86] Article 22 on the restriction of automated decision-making,[87] and Articles 13-15,[88] to be read with Article 22.[89] As AI machines are built and trained on data, Article 22 provides some

[79] Emil Protalinksi, 'ProBeat: Why Google is really calling for AI regulation' (Venture Beat Blog, 24th January 2020) <https://venturebeat.com/2020/01/24/probeat-why-google-is-really-calling-for-ai-regulation/> accessed 19th May 2020.

[80] United Nations, *Promotion and protection of the right to freedom of opinion and expression* (August 2018, General Assembly, Seventy-Third Session) 15.

[81] ibid 20.

[82] Kalliopi Spyridaki, 'GDPR and AI: Friends, Foes or something in between?' (SAS Insights, Data Management, ND) <https://www.sas.com/en_gb/insights/articles/data-management/gdpr-and-ai--friends--foes-or-something-in-between-.html#/> accessed 21st July 2020.

[83] Eric Winston, 'GDPR – How does it impact AI?' (Information Age Blog, Data Protection and Privacy, 5th June 2023) <https://www.information-age.com/gdpr-impact-ai-123483399/> accessed 21st July 2023.

[84] Centre for Information Policy Leadership, *Artificial Intelligence and Data Protection – How the GDPR Regulates AI* (2020) 19.

[85] Daniel Castro and Eline Chivot, 'Want Europe to have the best AI? Reform the GDPR' (Privacy Perspectives, 23rd May 2019) <https://iapp.org/news/a/want-europe-to-have-the-best-ai-reform-the-gdpr/> accessed 12th May 2020.

[86] GDPR (n 1) Article 4(4).

[87] ibid Article 22.

[88] ibid Article 13-15.

[89] ibid Article 22.

restraint and consideration for all industries that wish to use automated means to help aid efficiency.[90] Other provisions which are of particular relevance to AI systems include the requirement for processing to be fair,[91] the principle of data minimisation,[92] data protection impact assessments,[93] and the right to an explanation.[94]

Although there is recognition of the relationship between AI and the GDPR, a recommended framework which uses the GDPR as inspiration for AI regulation is currently lacking in the literature. In terms of the scope of the thesis focusing on AI that processes personal data, the relationship with the GDPR should be used as a benefit, and as a starting point and basis for a suggested new regulation. However, it must be recognised that whilst the GDPR covers aspects such as fairness and the right to an explanation, it fails to fully address every challenge posed by AI systems. For example, developments in ML presents novel challenges that the current provisions under the GDPR may not fully anticipate.

Whilst the GDPR provides a strong foundation and building block for regulation, it is not the complete remedy for the regulation of AI. Future regulation must be developed with use of the GDPR as inspiration and as a starting point, ensuring that its strengths are built upon whilst the limitations and loopholes are sufficiently addressed when it comes to regulating AI.

1.1.6 Impact of Persuasive Sources

Despite no clear, specific regulation for AI systems in England and Wales, non-binding proposals and reports have been introduced domestically, for example, the UK Government’s White Paper on AI,[95] the Information Commissioners Office’s (ICO) Auditing Framework,[96] and the recently introduced Artificial Intelligence

[90] Winston (n 85).
[91] GDPR (n 1) Article 5(1)(a).
[92] ibid Article 5(1)(c).
[93] ibid Article 35.
[94] ibid Articles 13-15 and Recital 71.
[95] Department for Science, Innovation and Technology, *White Paper: A pro-innovation approach to AI regulation* (March 2023, updated 2024).
[96] Information Commissioner’s Office, *Guidance on the AI auditing framework* (2020).

(Regulation) Bill.[97] Growing research from the EU has also allowed the development of the Guidelines for Trustworthy AI,[98] the recently approved AI Act,[99] and the proposed AI Liability Directive.[100]

The EU's non-binding Guidelines for Trustworthy AI were released in 2019, and promoted AI in compliance with the law and ethics, by ensuring adherence to ethical principles and values.[101] The Guidelines also intended to put policymakers on the right track for future regulation.[102] Chivot, Stolton, and Meyer praised the Guidelines for the EU's swift release of ethical rules.[103] However, the Guidelines were subject to much criticism, due to the lack of detail regarding the future negative impacts of AI, leading to many, including Stolton and Meyer, having concerns about the future unforeseen high-impact ramifications of AI usage.[104] Concern was also shown in the Guidelines tone and word choice, which seemed to suggest AI as inherently untrustworthy, which fails to promote public trust.[105] Meyer also stated that the EU's focus on ethics primarily, in comparison to value and accuracy is a losing strategy,[106] and although it is progress in the general area of AI, just ethics will never be enough.[107] It is inevitable that some machines will perform poorly, and regulation is needed soon to ensure the legal and social compatibility between AI and humans.[108]

[97] Artificial Intelligence (Regulation) Bill, HL Bill (2023-24).

[98] High-Level Expert Group on Artificial Intelligence, *Guidelines for Trustworthy AI* (Independent group set up by the European Commission, 8th April 2019).

[99] European Parliament, *Provisional Agreement Resulting from Interinstitutional Negotiations – Proposal for a regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts 2021/0106 (COD)* (2019-2024).

[100] European Commission, *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to Artificial Intelligence (AI Liability Directive)*, COM (2022).

[101] High-Level Expert Group on Artificial Intelligence (n 100) 2.

[102] Ibid, 2.

[103] Eline Chivot, 'Relying on competitive advantage of AI Ethics is a losing strategy for Europe' (Center for Data Innovation Press Release, 8th April 2019) < https://datainnovation.org/2019/04/relying-on-competitive-advantage-of-ai-ethics-is-a-losing-strategy-for-europe/> accessed 18th July 2021; Samuel Stolton, 'Artificial intelligence presents 'black swan' ethical issues, Commission report says' (Euractive Digital Blog, AI news, 8th April 2019) <https://www.euractiv.com/section/digital/news/artificial-intelligence-presents-black-swan-ethical-issues-commission-report-says/> accessed 9th April 2019; David Meyer, 'Europe thinks ethics is the key to winning the AI race. Not everyone is convinced' (Fortune Blog, 8th April 2019) <http://fortune.com/2019/04/08/eu-ai-ethics-principles/> accessed 10th April 2019.

[104] Stolton (n 105); Meyer (n 105).

[105] Chivot (n 105).

[106] Meyer (n 105).

[107] Stolton (n 105).

[108] Lee Gluyas and Stefanie Day, 'Artificial Intelligence- who is liable when AI fails to perform?' (CMS Blog, Law and Tax, 2018) <https://cms.law/en/GBR/Publication/Artificial-Intelligence-Who-is-liable-when-AI-fails-to-perform> accessed 24th March 2019.

The EU's draft and White Paper on AI regulation covered the European approach of excellence and trust towards AI systems, by providing key pillars of a regulatory framework for AI.[109] The EC's approach is based on the fundamental values of the EU, including the respect for human dignity, pluralism, non-discrimination and protection of privacy, and the report showed a willingness to cooperate at an international level if the suggested approach respects these values.[110] The EC acknowledged a range of issues that need to be considered in relation to AI regulation, including flaws in the technology, the training datasets, and data quality issues.[111] These risks materialising make the characteristics of AI systems increasingly difficult to fit into current legislation, hence, making it difficult for appropriate remedies to be available under current law.

Although the EC addressed several issues related to AI, including safety, liability, and transparency, and set out a suggested and favoured proposal in regard to a framework, in some areas clarity was still lacking. The European Data Protection Supervisor (EDPS) also highlighted the importance that AI must have an agreed, standardised definition for future policy-making initiatives, however, they expressed regret that the White Paper presented more than one definition of AI, adding to the ambiguity of the matter.[112] For regulation to be effective, it must clearly define what it being regulated, and unfortunately, there exists countless definitions of AI amongst experts in the field,[113] posing a challenge that needs to be resolved. The EDPS advocates for a precautionary approach to AI regulation, in reflection of the approach taken by the EU concerning the GDPR.[114] The EDPS argues that the White Paper supports a 'risk-based approach' (an approach that came to fruition in the form of the AI Act proposal)[115] and expressed concern that this would hold the risk of limiting the

[109] European Commission, *Structure for the White Paper on Artificial Intelligence – a European Approach* (2020, Draft 12/12) 1
[110] European Commission, *White Paper on Artificial Intelligence – A European approach to excellence and trust,* COM (2020) 9.
[111] ibid 12.
[112] European Data Protection Supervisor, *EDPS Opinion on the European Commission's White Paper on Artificial Intelligence – A European approach to excellence and trust* (2020) 6.
[113] Scherer (n 70).
[114] European Data Protection Supervisor (n 114) 10.
[115] European Commission, *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts*, COM(2021).

applicability of the obligations, and that the measures suggested only address a portion of AI machines, predominantly those that are of 'high risk'.[116]

The ICO's framework for Auditing AI covers two main components; the government and accountability component, which discusses the measures organisations should have in place concerning data protection requirements, and the AI-specific risk component, which focuses on the potential data protection risks and the adequate risk management practices to be used.[117] The framework lists the following AI-specific risk areas:

- "*fairness and transparency in profiling;*
- *accuracy;*
- *fully automated decision-making models;*
- *security and cyber-security;*
- *trade-offs;*
- *data minimisation and purpose limitation; and*
- *the exercise of rights and impact on broader public rights*".[118]

The guidance identifies the technical means by which discrimination can be mitigated, and considers the processing of data which includes protected characteristics under the Equality Act[119] to assess and address discrimination in AI systems.[120] Kazim and Koshiyama express that more clarification is needed in the technical explanations of bias and discrimination.[121] They also highlight that although measures to assess and mitigate bias are discussed, the report later states that such metrics conflict with one another,[122] which lacks clarity. Also, the guidance fails to explore the issue of transparency at the same standard that lawfulness, bias, and discrimination are assessed, creating a considerable gap within the report.

[116] European Data Protection Supervisor (n 114) 13.
[117] Information Commissioner's Office (n 98) 4.
[118] ibid 36.
[119] Equality Act 2010.
[120] Information Commissioner's Office (n 98) 56.
[121] Emre Kazim and Adriano Koshiyama, 'A review of the ICO's Draft Guidance on the AI Auditing Framework' (SSRN, 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3599226> accessed 23rd July 2022.
[122] ibid; Information Commissioner's Office (n 98) 55-56.

1.1.7 Human Rights Impact

Human rights should be considered as a central concept to underpin future AI regulation. The UN states that human rights are inherent to human beings regardless of race, culture, ethnicity or status,[123] and have campaigned to ensure that human right standards are upheld worldwide. When considering the human right implications of AI, it is important to note the conceptions of human rights that vary across different legal frameworks. The UDHR for example, provides a broad, universalist framework which has heavily influenced international law and policy,[124] whereas regional instruments offer more focused perspectives on human right protection, relevant to the social and cultural status of where the instrument belongs.[125] With the rapid growth of AI, it is more important than ever that human rights protection is at the forefront of discussions. However, given the variations to the interpretation of human rights, it must be confirmed that for the purposes of this thesis, the argument toward human right protections stems from the Convention rights included within the ECHR and HRA,[126] which are seen as the cornerstone to UK constitution and the jurisdiction of England and Wales.

The UN has examined the impact AI will have on human rights, in particular the freedom of expression.[127] The importance of monitoring state and public sector use is highlighted, to ensure that human rights principles are complied with, and state that public consultations and human rights impact assessments should be completed.[128] As well as this, they share that states should ensure "*human rights are central to private sector design, deployment and implementation*",[129] arguing that existing regulations must be updated, especially data-protection regulation, and that regulatory schemes must be introduced to ensure the undertaking of impact assessments, and provide effective external accountability mechanisms.[130] The UN also acknowledge the importance of transparency, and express that companies

[123] United Nations, 'Human Rights' (UN Website) <https://www.un.org/en/global-issues/human-rights> accessed 4th September 2024.
[124] UDHR (n 4) ; Jack Donnelly, *Universal Human Rights in Theory and Practice* (Cornell University Press, 2013) 40.
[125] ECHR (n 3) ; Human Rights Act 1998.
[126] ibid.
[127] United Nations (n 82).
[128] ibid 10.
[129] ibid 21.
[130] ibid 21.

should make it clear where and how AI technology is used on platforms, services and applications.[131] This means that individuals are aware when they are subject to AI-driven decision-making, or if it plays a role in processing their data, which is important not only for transparency, but also to give users the necessary notice to understand and address the impact of AI technology on their human rights.[132]

AI systems have the capability to negatively interfere with and impact several human rights, including but not limited to Articles 8, 10 and 14 of the ECHR.[133] AI also affects fundamental rights such as due process, freedom of assembly, right to a life, work, and education to name a few, with the potential of having the reach to affect all human rights and freedoms.[134] Regulation on the use of AI needs to safeguard the protection of these rights, whilst also adhering to data protection principles.[135] However, this is currently lacking, causing concern in light of the growing complexity of AI technology.

An area of developing AI use is through the processing of natural language and recognition of speech, as well as the generation of it, leading to the successful sales and use of products such as Apple’s Siri and Amazon’s Alexa.[136] However, concerns have grown in regard to privacy following an increasing amount of reports that consumers were being listened to by more than just the machine.[137] Within the area of content moderation, easily accepted as a necessary mechanism in society, the question arises of where to draw the line between too much and too little moderation. It is argued that AI should take a precautionary approach in relation to content moderation, however, removing content which is not universally agreed to be harmful could result not only in a damage to reputation, but also could undermine user’s freedom of expression,[138] a discussion that encountered conflict in the debates prior to enactment of the Online Safety Act.[139] The UN acknowledge that particular attention should be given to the impact of AI on ethnic and religious minorities,

[131] ibid 22.
[132] ibid 22.
[133] ECHR (n 3) Articles 8, 10 and 14.
[134] ECHR (n 3) Articles 6, 11, 2, 23 and Article 2 of Protocol 1; European Union Agency for Fundamental Rights, *Facial Recognition Technology: Fundamental rights considerations in the context of law enforcement* (2019) 23.
[135] United Nations (n 82) 20.
[136] Jaimes (n 28).
[137] Alex Hern, ‘Amazon staff listen to customers’ Alexa recordings, report says’ *The Guardian* (11th April 2019).
[138] Cambridge Consultants (n 4) 5; ECHR (n 3) Article 10.
[139] Online Safety Act 2023.

political oppositions and activists, to aid in the prevention of bias, 'fake news', and the protection of human rights in general.[140] They suggest that state deployment of AI systems in particular should be subject to regular testing and monitoring by external and independent experts,[141] which would place a responsibility on states to ensure this could be achieved sufficiently.

In regard to FRT, Access Now reflect the view that mass surveillance constitutes a significant violation of fundamental rights and freedoms,[142] and is part of a body of over 40 organisations who have called on EU institutions to ensure biometric technologies that enable mass surveillance are indefinitely banned, both in law and practice.[143] Such technologies are suggested to infringe on fundamental rights relating to privacy, data protection, equality, freedom of expression, freedom of assembly, due process, and more, and are argued to not be 'prescribed by law', nor 'necessary' or 'proportionate' for the area of surveillance.[144] This has been proven in *R(Bridges)*, where the use of FRT was held not to be prescribed by law,[145] and therefore causing an unlawful infringement of Article 8 of the ECHR.[146] It should be highlighted however, that the ground relating to proportionality was rejected,[147] commenting that the balancing exercised when considering proportionality is not a mathematical one, but one that calls for judgement.[148]

This case highlights the need for regulation of AI systems, and identifies the limits and restrictions needed for technology that poses a serious infringement to rights. A series of tests on major technologies have found that algorithms particularly struggle in identifying darker-skinned females,[149] posing a significant risk to Article 14.[150] The American Civil Liberties Union report on Amazon's facial recognition system found a similar result, whereby the system incorrectly flagged those with darker skin more

[140] United Nations (n 82) 12.
[141] ibid 21.
[142] Access Now, *Submission to the Consultation on the 'White Paper on Artificial Intelligence – a European approach to excellence and trust'* (May 2020) 5.
[143] EDRi, *Ban Biometric Mass Surveillance* (2020).
[144] Access Now (n 140) 6.
[145] *R(Bridges) v CC of South Wales* (n 49) [58].
[146] ECHR (n 3), Article 8.
[147] *R(Bridges) v CC of South Wales* (n 49) [144].
[148] ibid [143].
[149] Ryan Daws, 'UK Police are concerned AI will lead to bias and over-reliance on automation' (Artificial Intelligence News Blog, 17th September 2019) <https://artificialintelligence-news.com/2019/09/17/uk-police-concerned-ai-bias-automation/> accessed 30th January 2020.
[150] ECHR (n 3) Article 14.

often.[151] AI bias occurs when an algorithm produces results that are systematically prejudiced, due to flawed assumptions in the ML progress, and results in a lack of public trust.[152] With the absence of transparency requirements, it may be impossible to identify potential AI bias, causing obvious concern for individual rights.[153] In agreement, the HLEG on AI express that it should be possible to demand a suitable explanation of the AI system's decision-making process in order to protect citizen's rights.[154] Berthelemy agrees, suggesting that human rights should guide the "*development, design and deployment and calls for enhanced transparency, disclosure obligations and robust data protection legislation, including effective means for remedy*",[155] in order to address the human rights concerns.

1.1.8 Bias, Transparency and Explainability

Bias

AI bias refers to systems which produce biased results that reflect and perpetuate human biases within society, including historical and current social inequality.[156] AI bias stems from two variations; data input where bias already exists, or a bias the algorithm classifies itself by highlighting the wrong data for the wrong purposes. AI making decisions based on biased data, or a biased algorithm, poses a significant risk to Article 14 of the ECHR,[157] highlighting the need for regulation to ensure such decisions do not go unnoticed. FRT and profiling, which is particularly widespread within state surveillance services,[158] arguably poses the most significant infringement to Article 14 of the ECHR[159], particularly to ethnic minorities, and specifically darker-skinned females.[160]

[151] Jacob Snow, 'Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots' (American Civil Liberties Union Blog, 26th July 2018) < https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28> accessed 4th February 2020.

[152] IBM, 'Trustworthy AI' (IBM website) <www.research.ibm.com/5-in-5/ai-and-bias/> accessed 11th June 2019.

[153] Centre for Information Policy Leadership (n 86) 13.

[154] High-Level Expert Group on Artificial Intelligence (n 100) 18.

[155] Chloe Berthelemy, 'UN Special Rapporteur Analyses AI's Impact on Human Rights' (European Digital Rights Blog, 7th November 2018) <https://edri.org/un-special-rapporteur-report-artificial-intelligence-impact-human-rights/> accessed 4th March 2020.

[156] IBM, 'Shedding light on AI bias with real world examples' (IBM website) <https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples> accessed 4th September 2024.

[157] ECHR (n 3) Article 14.

[158] ECHR (n 3) Article 14; Daws (n 147).

[159] ECHR (n 3) Article 14.

[160] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' [2018] 81 *Proceedings of Machine Learning Research* 1.

In terms of societal impact, it is imperative that the deployment, use and existence of AI is human-centric, and it must be ensured that this applies to all human beings.[161] Already in the United States (US), San Francisco became the first State to ban FRT used by law enforcement, due to concerns not only of the inaccuracy of the technology, but in the long national history of politicised and racially-biased state surveillance.[162] The ICO, backed by several pressure groups recently threatened legal action on the use of FRT,[163] calling it intrusive and demanding answers on transparency, accuracy, bias, effectiveness and the lack of national coordination.[164] In the US, Axon (the largest manufacturer of police body cameras) rejected the possibility of selling FRT after advice from an independent ethics board, which was created after the company had acquired two AI companies.[165] Evidently, the lack of regulation on this technology has led to a banning in some locations, and a pursuance in others.[166]

The over-dependence or disproportionate use of technology reflects the potentially devastating impact AI could have on society if the technology is rolled out too early, or without proper means for consistent and repeated testing and requirements to act if signs of bias begin to show. Looking at the UK, there have been many cases which highlight claims of police racially profiling, especially through the conduct of stop and search, and throughout the COVID-19 lockdown.[167] It is imperative that AI technology does not contribute to an already highlighted issue of racial and/or other inequities, proving the need for regulation to address these matters.[168]

[161] Centre for Information Policy Leadership (n 86) 17.
[162] Veena Dubal, 'San Francisco was right to ban facial recognition. Surveillance is a real danger' *The Guardian* (30th May 2019).
[163] Liberty, 'Legal Action' (Pressure Group Website) <https://www.libertyhumanrights.org.uk/?s=legal+action> accessed 3rd January 2022; Big Brother Watch, 'Active Campaigns' (Pressure Group Website) <https://bigbrotherwatch.org.uk/campaigns/ accessed 3rd January 2022.
[164] Lizzie Dearden, 'Information Commissioner threatens legal action against police using 'dangerous and inaccurate' facial recognition technology' *The Independent* (15th May 2018).
[165] Axon AI and Policing Technology Ethics Board, *First Report of the Axon AI & Policing Technology Ethics Board* (2019) 9.
[166] Jim Gill, 'Where Does eDiscovery Fit in the Facial Recognition Conversation?' (JD Supra Blog, 5th August 2019) <https://www.jdsupra.com/legalnews/where-does-ediscovery-fit-in-the-facial-40933/> accessed 14th June 2020.
[167] Vikram Dodd, 'Cases that highlight claims of police racial profiling in England' *The Guardian* (9th July 2020, London); Ben Quinn and Frances Perraudin, 'London police accused of racial profiling in lockdown searches' *The Guardian* (16th May 2020, London).
[168] Axon AI and Policing Technology Ethics Board (n 162) 26.

The Centre for Data Ethics and Innovation's review into bias in algorithmic decision-making[169] acknowledges the challenges in addressing direct and indirect bias, and defining 'fairness'.[170] The review addresses the use of AI tools that carry significant ethical risks in relation to bias, and highlight the possibility that historical data would most likely be used in the AI training, which would include significant historical bias.[171] In relation to this, law enforcement in the UK have expressed concern that the use of AI in their operations may lead to increased bias and an over-reliance on automation.[172] The Centre for Data Ethics and Innovation found that a majority felt the use of AI may amplify prejudices.[173] This concern stems from the current and on-going issue of racial profiling within the police force,[174] and the likelihood of these prejudices making their way into AI algorithms if they are trained on existing police data.[175] To address these issues, the Centre for Data Ethics and Innovation believe that new technologies should be trialled in a "*controlled way prior to implementation, to establish whether or not a certain tool is likely to improve the effectiveness of a policing function*".[176] However, they also acknowledge that whilst public scrutiny of the tool is growing, currently, clear guidance for the ways in which police should conduct such trials and deploy AI technology is lacking, which creates the risk of mass adoption without proper consideration of the ethical concern of racial bias.[177]

Transparency

AI transparency typically refers to the ability to 'see inside' AI technology, with various stakeholders being able to understand the processes and decisions made by AI systems, and is seen as vital for building trust in systems.[178] The issue of AI transparency has been raised, namely in how to explain and understand the workings and results of an algorithm, issues relating to copyright infringement, and the issue of too much transparency, allowing individuals having the ability to

[169] Centre for Data Ethics and Innovation, *Interim Report: Review into bias in algorithmic decision-making* (2019).
[170] ibid 4.2; Equality Act 2010 s4.
[171] Centre for Data Ethics and Innovation (n 166) 15.
[172] Daws (n 147).
[173] Centre for Data Ethics and Innovation (n 166) 6.
[174] Equality and Human Rights Commission, *Stop and think – A critical review of the use of stop and search powers in England and Wales* (2010) 6
[175] Daws (n 147).
[176] Centre for Data Ethics and Innovation (n 166) 16.
[177] ibid 16.
[178] Stefan Larsson and Fredrik Heintz, 'Transparency in artificial intelligence' [2020] 9(2) *Internet Policy Review* 1.

manipulate or game an AI system.[179] The Science and Technology Committee highlights the importance of being able to 'inspect' algorithms so when cases arise where AI has gone wrong, the logic behind the decision made could be investigated.[180] Marsden and Nicholls express that AI is currently being used with "*little to no transparency*",[181] from the public use of FRT, to the use of content moderation to remove 'fake news' from social media platforms.[182] Yet currently, consumers have no disclaimer to make them aware of these technologies, nor to any remedy if their rights are potentially infringed, and therefore an interoperability remedy is advocated for which has the ability to allow regulators see within AI's 'black box'.[183]

Ofcom acknowledge that due to the public apprehension towards AI, systems are held at a higher standard in comparison to human moderators, and therefore, developing transparent and explainable AI is increasingly important.[184] Several AI learning techniques stem from models to support ANNs, which although designed to replicate the way the human brain learns, AI developers are often unable to understand or explain the algorithm outputs, creating a 'black box' whose inner workings are hidden, leading to considerable concern when deploying such systems into the real world. Reed expresses however that the more transparency mandated, the less AI can improve their use through ML, as full transparency would limit systems from evolving, and would require them to capture and upload their dataset before an improved version is approved, creating a delay.[185] This would prevent AI not only from progressing, but from performing to the best of its abilities.

[179] The Alan Turing Institute, 'A right to explanation' (Advice from Turing researchers) <https://www.turing.ac.uk/research/impact-stories/a-right-to-explanation> accessed 4th March 2020.

[180] Science and Technology Committee, *Robotics and Artificial Intelligence* (Fifth Report of Session 2016-2017, House of Commons, 2017) 18.

[181] Chris Marsden and Rob Nicholls, 'Interoperability: A solution to regulating AI and social media platforms' (Tech Law for Everyone Blog, ND) <https://www.scl.org/articles/10662-interoperability-a-solution-to-regulating-ai-and-social-media-platforms> accessed 5th March 2020.

[182] ibid.

[183] Lilian Edwards and Michael Veale, 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?' [2018] 16(3) *IEEE Security and Privacy* 46; Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" is probably not the remedy you are looking for' [2017] 16 *Duke Law and Technology Review* 18; Lilian Edwards and Michael Veale, 'Clarity, surprises, and further questions in the Article 29 Working Part draft guidance on automated decision-making and profiling' [2018] 34(2) *Computer Law and Security Review* 398.

[184] Cambridge Consultants (n 4) 45.

[185] Chris Reed, 'How Should we regulate Artificial Intelligence?' [2018] 376(2128*) Philos Trans A Math Phys Eng Sci* 13.

Also, it is difficult to determine the fundamental principles which could offer the basis of transparency that should be imposed, and more importantly, what kind of transparency and on what type of AI systems.[186] Reed argues that in some cases, such as an AI system producing an overall benefit to society, and where the loss to individuals is minor, it should be legally acceptable for a complete lack of transparency.[187] However, in cases where human rights or a significant impact to an individual is concerned, more transparency should be demanded. Transparency can be categorised into possessing a prospective element, which means individuals must be informed about the ongoing data processing before it takes place, and a retrospective element, meaning the ability to trace back how and why a particular decision was reached.[188] The prospective element can be produced in the form of transparency, whereas the retrospective element can be in the form of explainability. Although complex, transparency can help mitigate issues of fairness, discrimination, and trust, all of which have been major concerns within the academic field,[189] and if regulators were to enforce provisions for an explanation and clarity about the data used, sufficient methods would be needed to inspect algorithms and their results.[190]

Explainability

The concept of AI explainability relates to AI systems having the ability to provide clear and understandable explanations for their decisions and actions.[191] In line with transparency, the ICO states that explainability is also an essential ingredient towards the responsible progression of AI and ML technologies.[192] Requiring an explanation of decisions is a significant safeguard to protect individual citizens from unfair or unethical AI.[193] To support the ethical, legal and technical governance of AI, Cath expresses that explainability and interpretability are possible mechanisms to

[186] ibid.
[187] ibid.
[188] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz and Aurelia Tamo-Larrieux, 'Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns' [2019] 6 *Big Data and Society* 1.
[189] Andrew Bert, 'The AI Transparency Paradox' (Harvard Business Review Blog, 13th December 2019) <https://hbr.org/2019/12/the-ai-transparency-paradox> accessed 4th March 2020.
[190] Virginia Dignum, 'Responsible Autonomy' (Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017) <https://arxiv.org/pdf/1706.02513.pdf> accessed 19th June 2023.
[191] European Data Protection Supervisor, *TechDispatch: Explainable Artificial Intelligence* (2023) 3.
[192] Information Commissioner's Office, *Project ExplAIn Interim Report* (2019) 6.
[193] ibid 6.

increase algorithmic fairness, transparency and accountability.[194] As AI is currently in use in the UK, the ICO highlights the urgency for the responsible design and implementation of explainable AI products and services.[195] Cath also poses the question of whether there exists, or should exist a ‘right to an explanation’ for algorithmic decisions,[196] in which an example of this can be found within Recital 71 of the GDPR.[197]

Copeland argues the right to an explanation within the GDPR provides established principles for the fair and responsible use of data within UK and EU legislation.[198] Data Ethics expresses that the right to explanation should be interpreted functionally, flexibly and at a minimum, should enable the data subject to exercise their rights under the GDPR and human rights legislation.[199] However, the extent in which the GDPR actually provides the right is heavily debated.[200] There is a concern shared in light of the fact that Recitals are not binding. Although originally included in drafting, the right is not included within the binding articles of the text, due to its removal during the legislative process.[201] Wachter, Mittelstadt and Floridi highlight the resistance to making a right to explanation legally binding, acknowledging the need to show decision-makers it is realistic.[202]

In light of this, Edward and Veale acknowledge that transparency in the form of a ‘right to an explanation’ being introduced would make a promising remedy towards the algorithmic ‘black box’ issue, and encourages the promotion of challenge, redress and accountability.[203] However, they argue that the law is “*restrictive,*

[194] Corinne Cath, ‘Governing Artificial Intelligence: Ethical, Legal, Technical Opportunities and Challenges’ [2018] 376(2133) *Philosophical Transactions of the Royal Society A.*
[195] Information Commissioner’s Office (n 187) 6.
[196] Cath (n 189).
[197] GDPR (n 1) Recital 71.
[198] Eddie Copeland, ‘Does the public sector really need a code of AI ethics? (Nesta Blogs- Government Innovation, 15th February 2019) <https://www.nesta.org.uk/blog/does-public-sector-really-need-code-ai-ethics/> accessed 15th March 2020.
[199] DataEthics, ‘GDPR Does Entail a Right to Explanation’ (Dataethics Blog, 2nd December 2017) <https://dataethics.eu/study-gdpr-entail-right-explanation/> accessed 4th March 2020.
[200] Bryce Goodman and Seth Flaxman, ‘European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation’ [2017] 38(3) *AI Magazine* 50; Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’ [2017] 7(2) *International Data Privacy Law* 76*;* Andrew Burt, ‘Is There a Right to Explanation for Machine Learning in the GDPR?’ (Privacy Tech Blog, 1st June 2017) <https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/> accessed 22nd December 2019.
[201] Wachter, Mittelstadt and Floridi (n 195).
[202] ibid.
[203] Edwards and Veale (n 179).

unclear and contradictory"[204] in relation to when an explanation can be demanded, and that the narrowly defined right within the GDPR is not compatible with how modern ML technologies are being developed.[205] Wachter, Mittelstadt and Floridi also consider that the right to an explanation is viewed as an "*ideal mechanism*"[206] to enhance the transparency and accountability of automated decision-making, but express that the GDPR lacks clear language, effecting the clarity of the rights and safeguards against automated decision-making, which needs to be improved.[207]

1.2 The Gaps in the Literature

One of the main issues concerning AI regulation is the lack of a universal definition. This could be due to AI being an umbrella term, which covers various types of systems, machines, and algorithms, which work on different levels of complexity. Following this, the thesis advocates for a clear definition of the scope of discussion, something that has been an issue to identify within the literature. A more focused definition reflects the view that AI cannot be regulated as a whole, which goes against some academic suggestions and proposals.[208] For regulation to be effective, all main terms need to be defined, and without a universal definition for AI, it would be inadequate to base a regulation on a term which is too complex or too broad to define.

On this basis, this project advocates for regulation specifically for AI, with stronger mechanisms for systems that process personal data. This distinction is easier to categorise and is more focused, and therefore regulation will have an improved ability to provide adequate protection to individuals whilst also promoting the ethical and technological progression of AI. Also, AI that processes personal data has a direct correlation to the GDPR, which provides further justification for regulation focused on this type of AI, as well as using the GDPR as a basis for a future framework. Therefore, for the purposes of this research, the term AI will take inspiration from the EU's AI Act,[209] and include products or components that have

[204] ibid.
[205] ibid.
[206] Wachter, Mittelstadt and Floridi (n 195).
[207] Wachter, Mittelstadt and Floridi (n 195).
[208] High-Level Expert Group on Artificial Intelligence (n 100); Scherer (n 70).
[209] Provisional Agreement for the AI Act (n 103) Article 3(1).

differing degrees of intelligence and autonomy, which receive data inputs to produce outputs, and can perform one or more specific human tasks. This focus allows for a sufficient exploration of the overall topic, as well as a scope which can be achieved within the word count of the thesis.

The academic literature is vast in the debate for AI regulation, with the majority agreeing that existing principles are not sufficient for the growing complexities of AI.[210] Although regulation is now approved in the EU,[211] this research intends to fill the gap regarding AI regulation in England and Wales that focuses predominantly on systems that process personal data. The research uses a combination of different solutions, whilst also incorporating a techno-legal-ethics perspectives. This drawing together of perspectives also adds a new view to literature, with the added advantage of learning from the AI Act's mistakes. The challenges which affect the various perspectives are discussed and resolved together, making solutions to regulation have a sufficient justification. This thesis also critically analyses Article 22 of the GDPR[212] to justify subsequent reform of the DPA, to ensure alignment between future regulation and data protection rules. This reflects the view that existing legislation and principles are not currently sufficient for the complexities of AI, which directly disagrees with the ICO, who have previously suggested existing data protection principles are adequate to cover the issues posed.[213] By highlighting the potential violation of human rights, this research will argue that more needs to be done to protect citizen rights.

1.3 Research Aims, Problem, Questions, and Hypothesis

1.3.1 Research Aims and Objectives

The research poses the overarching question of 'to what extent should AI be regulated to ensure an ethical framework centred on human rights?'. The research has the following overall aim:

[210] Mark MacCarthy, *AI needs more regulation, not less* (on behalf of The Brookings Institution's Artificial Intelligence and Emerging Technology initiative, 2020); Mark Coeckelbergh, *AI Ethics* (The MIT Press Essential Knowledge Series, 2020) 145; Jacob Turner, *Robot Rules – Regulating Artificial Intelligence* (Palgrave Macmillan, 2019) 36.

[211] Provisional Agreement for the AI Act (n 103).

[212] GDPR (n 1) Article 22.

[213] Information Commissioner's Office, *Guidance on AI and Data Protection* (2020, updated 2023) 2.

(Research Aim 1) To produce up-to-date, justified and well-researched solutions concerning human rights based regulation for AI, using the GDPR as a basis for a new framework.

To achieve this aim, the thesis will address the following objectives:

(Research Objective 1) *To critically evaluate the ethical, legal, and technological issues posed by AI, to propose how regulation would best address these challenges.*

(Research Objective 2) *To assess the human right implications caused by AI, to develop new regulatory safeguards that effectively protect society and their rights.*

(Research Objective 3) *To examine current liability regimes and demonstrate the legal uncertainties they face with AI, to propose solutions to address this lack of clarity.*

(Research Objective 4) *To critically analyse the impact on data protection through the lens of the GDPR, and to propose potential areas of reform to strengthen alignment between AI systems and data protection rules.*

To justify the need for a framework on AI that processes personal data, the uncertainty and lack of clarity within the current legal landscape will be highlighted. The solutions produced aim not only to be realistic for use in society, but also ethical, conforming to the DPA and the ECHR.

1.3.2 Research Problem

For AI to be accepted and successfully coexist within society, redress mechanisms are crucial, and rules are considered a necessity to provide monetary security.[214] The complexities of AI have revealed a gap in the current liability[215] and causation rules,[216] which are too out-dated for the era of the technological advances AI offers.

[214] Caroline Cauffman, ‘Robo-liability: The European Union in search of the best way to deal with liability for damage caused by artificial intelligence’ [2018] 25(5) *Maastricht Journal of European and Comparative Law* 527.
[215] *R v White* [1910] 2 KB 124.
[216] *R v Dalloway* [1847] 2 Cox 273.

Considering incidents where AI machines are found at fault, existing laws are difficult, if not impossible to apply, highlighting the need for a new framework based on human and machine interaction.[217] Currently, with the absence of legislation specifically concerning AI in England and Wales, compensation to victims for injury or damage from a failure of AI would most likely be carried out under the tort of negligence.

However, this could be extremely problematic due to the claimant needing to establish that a duty of care was owed, that it was breached, and that breach was the cause of the damage suffered.[218] In the event of AI causing damage, and those actions of the machine considered unforeseeable, the claim would fail, resulting in no one being liable, which would cause several problems.[219] Also, advancements in AI means that soon enough, machines will be fully autonomous, perhaps removing the need for human intervention altogether, which creates further complexity when establishing both proximity and foreseeability for liability.[220] A new framework also needs to address the ethical issues of AI, especially in regard to the protection of human rights and data protection, to ensure the safe, fair and useful progression of AI, which if developed and regulated in the correct way, could successfully transform several industries.

1.3.3 Research Questions

The thesis is based on the following research questions:

Research Question 1 (RQ1): *How do AI systems challenge ethics, human rights, and current liability rules, and how would a new framework address these issues?*

The research intends to make a full consideration of the current ethical, legal, and technological issues surrounding AI, including bias, the lack of transparency and explainability, and the human rights impact. The research analyses the impact made

[217] Richard Kelley, ‘Liability in Robotics: An International Perspective on Robots as Animals’ [2010] 24(13) *Advanced Robotics* 1861.
[218] *Donoghue v Stevenson* (n 59).
[219] Gluyas and Day (n 110).
[220] Kowert (n 71).

to Articles 8, 10 and 14 of the ECHR[221] through the use and deployment of AI. In addition, an exploration is made concerning how these issues could be addressed with the introduction of new legislation in England and Wales.

Research Question 2 (RQ2): *To what extent does the GDPR fail to account for the unique challenges posed by AI systems, and how could reform of the DPA drive more ethical AI use?*

The thesis highlights the importance of the GDPR in relation to AI that processes personal data. The current impact made to the use of AI through the GDPR is analysed, whereby specific Articles are linked with previously discussed issues of transparency and explainability.[222] The research also proposes areas of reform within the DPA,[223] whilst highlighting the potential areas of use withtin data protection regulation for the basis of AI regulation.

Research Question 3 (RQ3): *What limitations exist in recent attempts to regulate AI made by the EU and UK, and in reflection, what solutions can be proposed to strengthen safeguards whilst encouraging ethical innovation?*

An exploration is made into the most recent developments and proposals for regulation, the benefits of the approaches, as well as highlighting sections that may need further amendments to sufficiently protect human rights and data protection, whilst also clarifying liability rules. The research analyses potential solutions that exist in the field of AI regulation and evaluates these against the thesis' recommendations of using the GDPR as a foundation for a new framework.

RQ1 intends to lay the foundations of the research, through identifying and analysing the core ethical, human rights and legislative issues posed by AI. This analysis is important as it reflects the key challenges that future regulation must address. This question sets the stage for RQ2, allowing a narrower focus of where current frameworks fall short, highlighting how a new framework could be beneficial in

[221] ECHR (n 3) Articles 8, 10 and 14.
[222] GDPR (n 1) Articles 13-15 and Recital 71.
[223] Data Protection Act 2018.

addressing the identified shortfalls. This allows a natural progression into the focus on the GDPR and data protection rules, which allows the research to focus on measures that have a significant impact on the deployment and use of AI. RQ3 allows for a consolidation of insights made from RQ1 and RQ2, to propose comprehensive solutions to AI regulation to strengthen safeguards and protection for human rights, whilst also fostering ethical innovation. The solutions proposed under RQ3 are informed by the ethical discussion made in answering RQ1, and take into consideration the findings of RQ2, to ensure the recommendations align with already enforced regulations.

1.3.4 Hypothesis

This research tests the hypothesis that using the GDPR as a foundation for a new framework will have a significantly positive effect on AI ethical and technological development, concerning the issues of transparency and explainability, and data protection. The research intends to align with the EC's ethical focus and respect for fundamental rights, to provide solutions towards a human rights centred framework to regulate AI in England and Wales. Domestic AI regulation is currently lacking, and for AI to progress in an increased safe, fair, and ethical manner, binding legislation needs to be introduced.

With increased deployment and use of AI available to the public and the state, clear and realistic rules are a necessity, therefore justifying the need to introduce solutions to address the ethical and legal issues, in the form of a new human rights centred AI framework. Also, in disagreement with previous comments from the ICO, this research argues that the existing data protection principles are insufficient, and that more can be done to protect citizens' rights.[224] As the scope of this project focuses predominantly on AI that processes personal data, using the GDPR as a basis for a new framework poses the best solution in ensuring the safe, fair, and ethical progression of AI.

[224] Information Commissioner's Office (n 208) 2.

1.4 The Importance of Research and the Contribution it Makes

This research contributes to the literature and adds to a relatively new subject area by providing original, up-to-date, and realistic solutions to solve the issues AI poses. Addressing the lack of domestic AI regulation is a necessity, particularly considering the EU's recent regulatory advances. This research draws together legal, ethical, and technological perspectives to focus on AI that processes personal data, to not only add to the existing literature on general AI regulation, but build to the areas of literature focused on AI and data protection.

This research demonstrates that reform of the DPA, in combination with a new framework that focuses on AI that processes personal data will have a substantially positive effect on the safe, fair, ethical, and technological progression of AI. The solutions produced comply with the standards within the GDPR and the ECHR, are well-justified through research, have an ethical and realistic focus, and include requirements, monitoring, and obligations in law. This thesis argues that existing principles of data protection are not enough to cover the increasing complexities of AI, and justifies why a new framework would produce an improved benefit in comparison to the current stance.

This project is innovative in the ways it seeks to understand the past and recent attempts at regulating AI. The project uses the themes of the GDPR and ECHR throughout, to enable the depth and exploration needed to answer the thesis' research questions. The GDPR already makes a significant impact on the use and training of AI, and therefore provides a sufficient basis for an AI-specific framework, which can be built upon in the future. Increasing the levels of transparency and explainability of machines would aid in correcting errors that exist in AI, to ensure protection to fundamental rights and safeguarding of data protection.

In the project, an ethical perspective will be sought through the human rights focus, especially regarding the issues of bias, transparency, and explainability. A technological view will be considered with respect to the impact the suggested regulation could make in terms of the progression and innovation of AI, and will be considered when suggesting limits or restrictions, to ensure they would work in practice and balance the interests of all parties. A legal focus will be made through a critical analysis of Article 22 of the GDPR,[225] and the examination of the most recent regulatory attempts in the area. This project is also current, providing valuable insight into a collection of recent literature, and worldwide developments concerning AI regulation and related data protection issues.

This research will also widen the understanding of AI machines and their impact, on primarily legal but also societal aspects, and will provide suggestions and recommendations for a new framework for AI, using the GDPR as a foundation. Rigorous safety standards and an established safety certification process for algorithms will be necessary as more AI systems are made available to the public.[226] To produce an appropriate framework in practice, input from AI experts would be necessary, due to the complexity of ML techniques and the general lack of understanding the public has towards AI.[227] In addition to this, advisory committees should be established to aid in deciding how AI systems should be regulated.[228] A framework should also be flexible enough to consider local and global considerations towards AI, for example, future national standards between countries or the need to comply with any future international treaties,[229] as well as future advancements of the technology.

1.5 Methodology

This research analyses academic material, proposals, recommendations, and current legislation, relying on both primary and secondary sources of law available from the public domain. Primary sources can either be mandatory, including case-law, legislation and regulations which are binding on the courts, or persuasive, whereby the courts can follow, but are under no obligation to do so. To understand and appreciate the current progression of AI, and the consequential effects from the current lack of regulation in England and Wales, primary and secondary legal

[225] GDPR (n 1) Article 22.
[226] F. Patrick Hubbard, 'Sophisticated Robots' Balancing Liability, Regulation, and Innovation' [2014] 66(5) *Florida Law Review* 1803.
[227] Woodrow Barfield, 'Liability for Autonomous and Artificially Intelligent Robots' [2018] 9(1) *Journal of Behavioural Robotics* 193.
[228] ibid.
[229] ibid.

sources need to be the focus and the basis of the approach to answer the research questions.

Given the interdisciplinary nature of AI, a strictly legal analysis would be inadequate in capturing the broader ethical and technological considerations which are integral to this research. AI raises novel issues that are relevant not only to the legal domain, but also ethical considerations such as fairness, transparency and accountability, and technological complexities, such as foreseeability, explainability and autonomy. The research therefore adopts a techno-legal-ethics approach, drawing together the three disciplines to fully validate and justify suggestions for AI regulation in England and Wales. This interdisciplinary perspective allows for a comprehensive view to be given, ensuring that the legal solutions proposed are grounded in ethical considerations and are technically feasible. The research includes a range of material on a subject which is current, and already attracts major attention, and aims to stand out in the inspiration taken from the GDPR. Linking AI and data protection offers a valuable and knowledgeable opinion, to be able to fill gaps in academic research and regulation material, produce suggestions towards a new framework, as well as make adequate suggestions for reform, strongly backed by an independent evaluation of academic opinion.

The most significant sources underpinning the legal focus of this research include the GDPR, which is currently a mandatory source due to the implemented DPA, the soon-to-be AI Act,[230] and the proposed AI Liability Directive,[231] which are now considered persuasive sources post-Brexit. Linking the legal and ethical perspectives, the ECHR[232] will also serve as a significant mandatory primary source. Secondary sources are always categorised as persuasive, as they are not binding on the courts; and a range of sources from legal, technological, and ethical disciplines will be analysed[233] to aid in justifying and supporting the argument, or to provide a contradicting opinion or different approach to the discussion.

[230] Provisional Agreement for the AI Act (n 103).
[231] Proposed for the AI Liability Directive (n 104).
[232] ECHR (n 3).
[233] European Commission (n 67); Cambridge Consultants (n 4); Centre for Information Policy Leadership (n 86); High-Level Expert Group on Artificial Intelligence (n 100); Information Commissioner's Office (n 98) ; Information Commissioner's Office (n 208).

In consideration of the rapid and significant development surrounding AI in terms of both technology and literature, the research intends to include secondary sources which were published most recently, and no longer than 10 years ago unless for a relevant and justified reason. This enables the discussion and research used to be up-to-date and current, as well as in line with any recent changes or developments in AI. This methodology enables the research produced to be modern, highlighting the most recent issues surrounding AI, and highlighting the importance of data protection, to propose justified and realistic solutions through drawing together, analysing and collecting research from a range of perspectives.

The techno-legal-ethics is a consistent perspective taken to underpin the discussion throughout the thesis. Considering the chapters in turn, Chapter Two intends to use this approach with a thematic analysis method, to identify and highlight the key themes in the literature, which will be used to set the doctrinal focus and discussion in future chapters. This method has been chosen due to the widespread issues that could be discussed relating to AI regulation, and the importance of maintaining an appropriate scope for this level of research. This method also helped to identify the key, recurring themes that link it to the overarching goal of the project, and each theme contributes to ensuring a sound regulatory approach to safeguard human rights.

Based on these identified themes, which include human rights, liability, data protection and regulation, Chapters Three, Four and Five utilise a doctrinal methodology that builds from the thematic analysis, to set the foundation and scope of discussion within the thesis. This research method is most appropriate due to the need for a thorough examination and analysis of current and future legislation, case-law and legal principles. The doctrinal approach allows the in-depth exploration of the legal instruments to identify gaps or inconsistencies, providing a solid foundation to assess the legal issues, whilst also allowing for the interdisciplinary perspective to play a part in the examination of the law. Through use of this method, the core research aim of producing well-researched, up-to-date, and well-rounded suggestions towards building a human rights centred framework for AI can be achieved.

Chapter Five also integrates some micro and macro comparative elements. The macro comparative approach allows for a comparison of the broader approaches and structures of proposals, allowing the research to draw on lessons from other jurisdictions. This analysis helps identify areas where the EU's AI Act may serve as a model, as well as areas where alternative approaches may be more effective in addressing specific concerns where the AI Act falls short. The micro comparative element, on the other hand, focuses on specific definitions and concepts across jurisdictions. This is crucial to understand how different legal systems address the key regulatory issues, enhancing the depth of research by highlighting best practices and potential pitfalls in AI regulation, providing a broader context for the recommendations proposed in the thesis.

By adopting an interdisciplinary approach and combining a thematic analysis, doctrinal methodology and comparative analysis, it is ensured that the research is comprehensive, well-structured, and capable of answering the RQs in a meaningful way. Together, these methods allow the research to present well-rounded, justified, and practical recommendations for creating a human rights centred framework for AI.

1.6 Layout

Moving forward, Chapter Two includes a literature review of the relevant primary and secondary sources of law, to identify the gaps within the literature. These gaps are linked to the research aims, problem and questions, hypothesis, contributions, and methodology of the thesis. The research materials include both mandatory and persuasive sources, to identify areas of concern and potential conflict which already exist in legislation and the literature. Current liability frameworks are used in the exploration of the current conflict with AI, and areas where clarity is lacking, or where further adjustment is necessary are also identified. The main persuasive sources reviewed in this Chapter include the EU's Guidelines for Trustworthy AI,[234] the White Paper on AI,[235] and the ICO's auditing framework.[236] In combination with a vast amount of academic literature, the existing and suggested recommendations in the

[234] High-Level Expert Group on Artificial Intelligence (n 100).
[235] European Commission (n 112).
[236] Information Commissioner's Office (n 98).

field will be evaluated, to justify the need for a new AI framework in England and Wales.

In Chapter Three, the issues and challenges of AI are explored, namely the difficulties in defining AI, the issues stemming from the lack of regulation, and the difficulties of placing liability. The gaps already identified in existing liability frameworks are explored and used to justify the need for a new framework for AI. Continuing, the human rights implications posed by the deployment and use of AI will be assessed, making consideration for the rights of predominant interference.[237] The potential conflict and challenges posed to Articles 8, 10 and 14 come into focus, arguably due to the highest likelihood of interference. These potential conflicts are highlighted to justify areas in which safeguarding needs to be stronger.

Chapter Four continues by assessing the current impact of the GDPR on AI systems that process personal data. This includes an in-depth analysis of the relationship between AI and the GDPR, to identify areas which lack clarity or are potentially conflicting or too complex to apply to AI. Relating to this, a critical examination is made towards Article 22 of the GDPR[238] to suggest potential areas of reform within the DPA to ensure the promotion of safe, fair, and ethical AI progression.

Following this, Chapter Five continues the evaluation of recommendations with a focus on the most recent proposals established by the EU, including the AI Act,[239] and proposed AI Liability Directive,[240] in addition to the recent UK proposals following Brexit. To assess the recommendations, this chapter will feature case studies to test how the proposals would work in practice. The thesis ends by consolidating and producing solutions to the presented argument, highlighting again the areas of reform recommended to the DPA, in combination with the introduction of a new human rights centred framework for AI, using alternative aspects of the GDPR as a basis.

[237] ECHR (n 3) Articles 8, 10 and 14

[238] GDPR (n 1) Article 22.

[239] Provisional Agreement for the AI Act (n 103).

[240] Proposed for the AI Liability Directive (n 104).

Chapter 2: Regulating AI: Key Themes and Legal Challenges

The purpose of this chapter is to explore the current stance of academic literature concerning AI, and the related research topics of human rights, liability, data protection, and regulation. This chapter provides a thematic analysis, to highlight and identify the key themes in the literature, and the gaps and conflicts within the current literature in relation to these key themes. By identifying these gaps, the chapter justifies the need for further research in this area, and situates the contribution of this research within the broader academic discourse. Although the academic research in this area covers a broad range of viewpoints and proposals, this review focuses on the themes that directly support the thesis' argument for new AI regulation. The chapter examines the current discussions in the literature surrounding AI's impact on human rights, and the challenges to liability and data protection frameworks, identifying where current analysis falls short. The literature raises important concerns, including the lack of transparency and risk of bias in systems, yet there is inadequate attention given to how frameworks could evolve to address these issues in a comprehensive way. This gap underpins the argument within this thesis for a future regulatory framework for AI, which will be explored in full in the later chapters.

Without the limitations of regulation, the industry-level development of AI technologies continues to be a contemporary and fast-paced area, and this is reflected throughout the literature.[241] Currently there is no mandatory regulation specifically focused on AI in England and Wales, however, the EU has made significant steps[242] with the recently approved AI Act,[243] being the first worldwide to issue a blanket regulation for AI technology. Before this, non-binding guidelines and White Papers were introduced,[244] however with no effect on the court system, concerns were raised towards the need for hard and compulsory legislation. The concern with non-binding regulation stems from the lack of enforcement, especially if the majority choose not to adhere, resulting in not only a waste of time and

[241] Boyang Chen, Zongxiao Wu and Ruoran Zhao, 'From fiction to fact: the growing role of generative AI in business and finance' [2023] 21(4) *Journal of Chinese Economic and Business Studies* 471; Dinesh Kalla, Nathan Smith, Sivaraju Kuraku and Fnu Samaah, 'Study and Analysis of Chat GPT and its Impact on Different Fields of Study' [2023] 8(3) *International Journal of Innovative Science and Research Technology* 827.
[242] European Commission (n 67); High-Level Expert Group on Artificial Intelligence (n 100); European Commission (n 112).
[243] Provisional Agreement for the AI Act (n 103).
[244] High-Level Expert Group on Artificial Intelligence (n 100); European Commission (n 112).

resources, but also concerns of non-regulated AI to continue.[245] It should be noted that this review focuses on the efforts and preliminary proposals and recommendations behind forming regulation for AI, in which discussion continues to address and analyse the most recently approved Act itself, in addition to other proposals worldwide within Chapter Five.

In reflection and to address RQ1, the literature review begins with an evaluation of the most substantial sources in relation to the topics of AI ethics and liability. Although this thesis touches upon ethical issues, particularly in regard to the human rights impact by AI, including bias, and the lack of transparency in systems, the ethical discussion is limited to areas that directly inform the legal and regulatory framework ideas proposed in this thesis. Broader ethical concerns, for example, such as the implications on autonomy and wider justice are beyond the primary scope of this thesis but are still recognised as essential to the wider discourse on AI regulation.[246] In agreement with the literature,[247] the thesis is written with the view that more needs to be done to protect human rights and that future regulation in England and Wales needs to be re-focused in respect of this, which is discussed further in Chapter Three, with an evaluation of the most recent regulatory developments in Chapter Five, and addressed with recommendations in respect of this to conclude the thesis in Chapter Six.

The review continues by emphasising the complexities in applying current liability regimes to AI systems. The EU's Liability for AI report,[248] introduced in 2019, acknowledges that AI poses new challenges to the area of liability due to its

[245] Castro (n 57) 9; Paula Klein, 'Rules for Robots: The Path to Effective AI Regulation' (MIT Digital Blog, 12th June 2019) <http://ide.mit.edu/news-blog/blog/rules-robots-path-effective-ai-regulation> accessed 21st February 2020.

[246] For a broader discussion on ethical issues in AI, see Carina Prunkl, 'Human Autonomy at Risk? An Analysis of the Challenges from AI' [2024] 34 *Minds and Machines* 26 ; Qian Hu, Yaobin Lu, Zhao Pan, Yeming Gong and Zhilin Yang, 'Can AI artifacts influence human cognition? The effects of artificial autonomy in intelligent personal assistants' [2021] 56 *International Journal of Information Management* 102250 ; Alessandra Buccella, '"AI for all" is a matter of social justice' [2023] 3 *AI and Ethics* 1143 ; Filippo Santoni de Sio, Txai Almeida and Jeroen van den Hoven, 'The future of work: freedom, justice and capital in the age of artificial intelligence' [2024] 27 *Critical Review of International Social and Political Philosophy* 659.

[247] Daniel Leufer and Ella Jakubowska, 'Attention EU regulators: we need more than AI 'ethics' to keep us safe' (Access Now Blog, 21st October 2020) <https://www.accessnow.org/eu-regulations-ai-ethics/> accessed 29th November 2020; Fanny Hidvegi and Daniel Leufer, 'Trust and excellence – the EU is missing the mark again on AI and human rights' (Access Now Blog, 11th June 2020) <https://www.accessnow.org/trust-and-excellence-the-eu-is-missing-the-mark-again-on-ai-and-human-rights/> accessed 29th November 2020; Rowena Rodrigues. 'Legal and human rights issues of AI: gaps, challenges and vulnerabilities' [2020] 4(3) *Journal of Responsible Technology* 1.

[248] European Commission (n 67).

complexity and self-learning capabilities through updates and operations, its limited predictability, and vulnerability to cyber threats, by which the report proposes recommendations on how these can be addressed.[249] The review continues by assessing the concerns with the current PLD[250] and the notion of negligence[251]. In review of these sources, it is evident within the literature that AI technology brings new issues and challenges to current liability frameworks, in particular the concepts of foreseeability and duties of care. Taking the view of the literature, Chapter Three explores further whether the current level of protection to developers and consumers from AI systems is sufficient under the UK liability frameworks. These findings are used to justify the idea of a new framework for AI that processes personal data, with recommendations proposed on this basis within Chapter Three and Chapter Six.

The second part of the review assesses the current stance of the literature of AI regarding data protection, to provide context of discussion surrounding RQ2. Data protection forms a significant part of this research, due to the scope of the research focusing predominantly on AI systems that process personal data. The literature in review of the GDPR is vast, receiving an arguably balanced amount of criticism and praise.[252] The GDPR was introduced to provide further protection to the right to the protection of personal data, however, the question can be posed whether the new complexities and challenges brought by AI systems are sufficiently covered under the legislation. Persuasive sources, including the Guidance on AI and Data Protection introduced by the ICO,[253] and other proposals and recommendations in relation to data protection and AI are also reviewed, to identify the parts of the GDPR that may be inadequate. In review of these sources, the thesis adopts the view that current data protection principles are not sufficient for AI, which is explored further, and related recommendations proposed in Chapter Four and Chapter Six.

To provide background discussion to RQ3, the final part of the review continues with an assessment of the challenges and options to regulation, and an evaluation of the

[249] ibid, 2-9.
[250] Product Liability Directive (n 66).
[251] *Donoghue v Stevenson* (n 59); *Caparo Industries plc v Dickman* (n 59).
[252] Privacy International and Article 19, *Privacy and Freedom of Expression in the Age of Artificial* Intelligence (April 2018) 23; Cambridge Consultants (n 4) 55; European Parliament, *The Impact of the General Data Protection Regulation (GDPR) on artificial intelligence* (European Parliamentary Research Service, June 2020) 3.
[253] Information Commissioner's Office (n 208).

preliminary proposals and efforts made in the steps leading up to the recently approved AI Act.[254] The persuasive reports of focus include the EU's Ethical Guidelines for Trustworthy AI,[255] released in 2019 as a form of non-binding regulation which aims to promote trustworthy AI, stating that AI must comply with three components; that they be lawful, ethical and robust.[256]

Following this, the review also provides an analysis of the preliminary White Paper on AI,[257] which was released in 2020 with an approach to excellence and trust. The purpose of the White Paper was to set out policy options to enable a trustworthy and secure development of AI in Europe, advocating for a consistent approach across the EU,[258] and was the first major step towards AI regulation in the EU. Also introduced in 2020, the ICO's Guidance on the AI Auditing Framework[259] will be reviewed to justify the need for AI regulation in England and Wales, and for a re-focused regulation on AI systems which process personal data. These matters are discussed and developed further in Chapter Five, which includes an analysis of the EU's AI Act and proposed AI Liability Directive,[260] in addition to the UK proposals,[261] and notable proposals from other jurisdictions.

2.1 AI Ethics and Liability

To provide background to RQ1: *What issues do AI systems pose to ethics, human rights, and current liability legislation, and how would a new framework address this?*

2.1.1 AI and Human Rights

The rise of AI systems has made a clear impact on the world around us, transforming sectors with the integration of systems that have new levels of complexity and autonomy. AI has made a positive impact and successful change in several industries, and the advantages of AI are discussed throughout the literature

[254] Provisional Agreement for the AI Act (n 103).
[255] High-Level Expert Group on Artificial Intelligence (n 100).
[256] ibid 2.
[257] European Commission (n 112).
[258] ibid 2.
[259] Information Commissioner's Office (n 98).
[260] Provisional Agreement for the AI Act (n 103); Proposed AI Liability Directive (n 104).
[261] Department for Science, Innovation and Technology (n 99).

a vast amount.[262] Within healthcare for example,[263] AI has been used to assist medical professionals in diagnosis and treatment, and to predict patients' health risks and outcomes.[264] Although AI has brought many successful changes, these are undermined in the literature by the growing concerns towards the challenges and potential negative impact this technology will bring to society, especially if left unregulated.[265] Academics have raised concerns in a vast number of areas of AI, including the potential for malicious use of AI,[266] the rise of autonomous weapons systems,[267] the risk of systems being hacked or dishonestly taken control of, AI behaving in unpredictable ways,[268] and the results of incorrect programming or misuse, especially within the public sector.[269]

Most of these concerns relate to the potential impact and effects AI technology has on individual citizens and their rights,[270] and due to AI increasingly being available for public use, and therefore affecting an increasing number of citizens, these concerns must be addressed to reduce the risk of pressure on the courts to respond to claims involving AI. In light of RQ1, this section begins by exploring the literature

[262] Cambridge Consultants (n 4) 8-9.

[263] Jagreet Kaur and Kulwinder Singh Mann, 'AI based HealthCare Platform for Real Time, Predictive and Prescriptive Analytics using Reactive Programming' [2017] 933(1) *Journal of Physics: Conference Series* 1; Sobia Hamid, 'The Opportunities and Risks of Artificial Intelligence in Medicine and Healthcare' (University of Cambridge, 2016) <https://api.repository.cam.ac.uk/server/api/core/bitstreams/d4b6cb45-f7fc-45bd-bcd2-679801cefbe0/content> accessed 15th August 2022; Jian Guan, 'Artificial Intelligence in Healthcare and Medicine: Promises, Ethical Challenges and Governance' [2019] 34(2) *Chinese Medical Sciences Journal* 76.

[264] Fei Jiang, Yong Jiang, Hui Zhi, Yu Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen and Yongjun Wan, 'Artificial intelligence in healthcare: past, present and future' [2017] 2 *Stroke and Vascular Neurology* 230; Abhimanyu Ahuja, 'The impact of artificial intelligence in medicine on the future role of the physician' [2019] 7 *Peer J* 1; Tamra Lysaght, Hannah Lim, Vicki Xafis and Kee Ngiam, 'AI-Assisted Decision-making in Healthcare' [2019] 11 *Asian Bioethics Review* 299; Ian Mundell, 'Healthcare AI' (Imperial College London, October 2019) <https://www.imperial.ac.uk/enterprise/long-reads/healthcare-ai/> accessed 12th November 2020.

[265] Jon Truby, 'Governing Artificial Intelligence to benefit the UN Sustainable Development Goals' [2020] 28(4) *Sustainable Development* 946; Amitai Etzioni and Oren Etzioni, 'Should Artificial Intelligence Be Regulated?' [2017] 33(4) *Issues in Science and Technology* 32 ; AI Now, *AI Now Report* (2019) 6-9.

[266] Future of Humanity Institute, University of Oxford, Centre for the Study for Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation and Open AI, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation* (2018) 6-7.

[267] Autonomous Weapons, 'Ban Lethal Autonomous Weapons' (access to videos and pledge) <https://autonomousweapons.org/> accessed 12th November 2020; Mary Wareham, 'Stopping Killer Robots' (Human Rights Watch, 10th August 2020) <https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and> accessed 12th November 2020.

[268] Cath (n 189).

[269] Bernd Wirtz, Jan Weyerer and Carolin Geyer, 'Artificial Intelligence in the Public Sector- Application and Challenges' [2018] 42(7) *International Journal of Public Administration* 596 ; Slava Mikhaylov, Marc Esteve and Averill Campion, 'Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration' [2018] 376(2128) *Philosophical Transaction of The Royal Society A* 1.

[270] Steven Livingstone and Mathias Risse, 'The Future Impact of Artificial Intelligence on Humans and Human Rights' [2019] 33(2) *Ethics and International Affairs* 141; Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy and Madhulika Srikumar, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' (Berkman Klein Center for Internet and Society at Harvard University, 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482> accessed 22nd July 2021.

in relation to human rights and AI to reflect the concerns that have been raised, followed by the reasons why additional safeguards are necessary for future regulation. Although only a few are directly related to AI,[271] the response and reflections of recent case-law make it clear that further regulation is needed to control AI technology and the new capabilities it brings. In light of these findings, Chapter Three discusses these issues further, including the importance of the ECHR when considering regulation for AI, which is used to produce and justify the recommendations set out in the concluding Chapter Six.

2.1.2 The Impact on Human Rights

The potential impact made to human rights by the use and development of AI technology has been discussed throughout the literature to a huge extent by academics and organisations alike. Several pressure groups and research projects are directly advocating for stronger safeguards to protect human rights in a broad sense,[272] including Access Now who have advocated for stronger and stricter legislation to address data protection, FRT, and behavioural prediction technologies,[273] and Amnesty International who express that concrete action needs to be taken rather than just principles.[274] Privacy International and Article 19 advocate for a larger and re-centred debate on the use of AI in "*human critical contexts*",[275] and Human Rights Pulse suggest that ethics not encoded in law is naïve.[276] Other sources in the literature are either focused on industry specific

[271] *R(Bridges) v CC of South Wales* (n 49); Kashmir Hill, 'Wrongfully Accused by an Algorithm' *The New York Times* (3rd August 2020) ; *Loomis v Wisconsin.* 137 S.Ct. 2290 (2017).

[272] Livingstone and Risse (n 265) ; Fjeld, Achten, Hilligoss, Nagy and Srikumar (n 265).

[273] Access Now, *The European Human Rights Agenda in the Digital Age (February 2020)* 3-8.

[274] Anna Bacciarelli, 'Ethical AI principles won't solve a human rights crisis' (Amnesty International, 21st June 2019) <https://www.amnesty.org/en/latest/research/2019/06/ethical-ai-principles-wont-solve-a-human-rights-crisis/> accessed 14th November 2020.

[275] Privacy International and Article 19 (n 247) 5.

[276] Aparajitha Narayanan, 'A Human Rights Framework is Necessary to Govern Artificial Intelligence' (Human Rights Pulse, 8th June 2020) <https://www.humanrightspulse.com/mastercontentblog/a-human-rights-framework-is-necessary-to-govern-artificial-intelligence> accessed 14th November 2020.

human right challenges, such as healthcare,[277] defence,[278] transport,[279] or focus on a specific issue, such as bias and unfairness,[280] privacy and data protection,[281] surveillance,[282] liability,[283] or accountability,[284] and the related implications to human rights. On the topic of regulation, whilst most of the literature concentrates on how AI needs to be developed and regulated to respect human rights,[285] the Human Rights, Big Data and Technology Project centre their research on whether fundamental human rights concepts and approaches need to be adapted in the era of technological advancement and big data,[286] which may be viewed by some as inappropriate.

[277] W. Nicholson Price II 'Artificial Intelligence in Health Care: Applications and Legal Implications' (University of Michigan, 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3078704> accessed 16th March 2020; Daniel Schönberger, 'Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications' [2019] 27(2) *International Journal of Law and Information Technology* 171; Michael Rigby, 'Ethical Dimensions of Using Artificial Intelligence in Health Care' [2019] 21(2) *AMA Journal of Ethics* 121.

[278] M. Cummings, 'Artificial Intelligence and the Future of Warfare' (International Security Department and US and the Americas Programme, 2017) <https://www.chathamhouse.org/sites/default/files/publications/research/2017-01-26-artificial-intelligence-future-warfare-cummings-final.pdf> accessed 20th March 2020; Peter Asaro, 'On Banning Autonomous Weapons Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making' [2012] 94 *International Review of the Red Cross* 687.

[279] European Parliament, *Artificial Intelligence in transport: Current and future developments, opportunities and challenges* (European Parliamentary Research Service, 2019).

[280] Future of Privacy Forum, *Unfairness by Algorithm: Distilling the harms of automated decision-making* (2017); David Danks and Alex London, 'Algorithmic Bias in Autonomous Systems' (Proceedings of the 26th International Joint Conference on Artificial Intelligence 2017) <https://www.researchgate.net/publication/318830422_Algorithmic_Bias_in_Autonomous_Systems> accessed 21st March 2020; Rachel Courtland, 'The bias detectives' [2018] 558 *Nature* 357; Mathias Risse, 'Human Rights and Artificial Intelligence: An Urgently Needed Agenda' [2019] 41(1) *Human Rights Quarterly* 1; Access Now, *Human Rights in the Age of Artificial Intelligence* (2018); Sahajveer Baweja and Swapnil Singh, 'Beginning of Artificial Intelligence, End of Human Rights' (LSE Blog, 16th June 2020) <https://blogs.lse.ac.uk/humanrights/2020/07/16/beginning-of-artificial-intelligence-end-of-human-rights/ > accessed 15th November 2020.

[281] Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' [2019] 2 *Columbia Business Law Review* 1; Dunja Mijatović , 'In the era of artificial intelligence: safeguarding human rights' (Open Democracy Blog, 3rd July 2018) <https://www.opendemocracy.net/en/digitaliberties/in-era-of-artificial-intelligence-safeguarding-human-rights/> accessed 15th November 2020.

[282] Steven Feldstein, 'The Global Expansion of AI Surveillance' (Carnegie Endowment for International Peace, 2019) <https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.pdf> accessed 27th July 2021; Antonio Aloisi and Elena Gramano, 'Artificial Intelligence is watching you at work. Digital Surveillance, Employee Monitoring, and Regulatory Issues in the EU Context' [2019] 41(1) *Comparative Labor Law and Policy Journal: Automation, Artificial Intelligence and Labour Protection* 95.

[283] David Vladeck, 'Machines without principals: Liability rules and artificial intelligence' [2014] 89 *Washington Law Review* 117; European Commission (n 67); J Kingston, 'Artificial Intelligence and Legal Liability' [2016] 33 *Research and Development in Intelligent Systems* 269.

[284] Han-Wei Liu, Ching-Fu Lin and Yu-Jie Chen, 'Beyond State v Loomis: artificial intelligence, government algorithmization and accountability [2019] 27(2) *International Journal of Law and Information Technology* 122; Hin-Yan Liu and Karolina Zawieska, 'From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence' [2020] 22 *Ethics and Information Technology* 321.

[285] Ben Hartwig, 'The Impact of Artificial Intelligence on Human Rights' (Dataversity, 8th May 2020) <https://www.dataversity.net/the-impact-of-artificial-intelligence-on-human-rights/> accessed 16th November 2020; Rodrigues (n 242).

[286] The Human Rights, Big Data, and Technology Project, 'About Us' <https://www.hrbdt.ac.uk/about-us/> accessed 17th November 2020.

2.1.3 'High Risk' AI Technology

In relation to specific rights, the literature is focused on the potential impact made to Article 8 and 14 of the ECHR from 'high risk' AI technology,[287] especially those technologies with capabilities of capturing biometric data, such as FRT. More focus is given to the specifics of profiling, whilst less consideration is given to artificial narrow intelligence (ANI) in the broader sense.[288] As highlighted by Human Rights Watch, the major areas of concern within the literature in relation to FRT are its use to monitor public spaces[289] and protests,[290] to track and profile minorities,[291] to flag suspects in criminal investigations,[292] and the concern of mass surveillance.[293] There exist several reports of companies refusing FRT contracts due to human right concerns, including Microsoft who refused to provide technology to California law enforcement officer's cars and body cameras.[294] Amazon have previously implemented a one-year ban on police use of their FRT to allow time for stronger ethical regulation to be in place, however they have allowed its use for more positive matters including to rescue human trafficking victims and find missing children.[295] Google and IBM have also previously withdrawn their FRT technology from sale,[296]

[287] Filippo Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz and Levin Kim, *Artificial Intelligence and Human Rights: Opportunities and Risks* (Berkman Klein Center Research Publication, 2018) 4; Laura Stănilă, 'Artificial Intelligence and Human Rights, A Challenging Approach on the issue of equality' [2018] 2 *Journal of Eastern European Criminal Law* 19; Molly Land and Jay Aronson, 'Human Rights and Technology: New Challenges for Justice and Accountability' [2020] 16 *Annual Review of Law and Social Science* 223.

[288] Privacy International and Article 19 (n 247) 19.

[289] Lizzie Dearden, 'Facial recognition becoming 'epidemic' in British public spaces' *The Independent* (16th August 2019); Antoaneta Roussi, 'Resisting the rise of facial recognition' [2020] 587 *Nature* 350.

[290] Nicola Habersetzer, 'Russian Activists Fights Use of Facial Recognition Technology' (Human Rights Watch, 18th October 2019) <https://www.hrw.org/news/2019/10/18/russian-activist-fights-use-facial-recognition-technology> accessed 17th November 2022; Paul Wallis, 'Op-Ed: Counting protestors with AI changes the game forever' (Digital Journal, Technology, 8th July 2019) <https://www.digitaljournal.com/tech-science/op-ed-counting-protestors-with-a-i-changes-the-game-forever/article/553608> accessed 17th November 2020.

[291] Paul Mozur, 'One Month, 500,000 Face Scans: How China is Using AI to Profile a Minority' *The Seattle Times* (14th April 2019); Clare Garvie, 'Garbage in, Garbage out' (Georgetown Law, Center on Privacy and Technology, 16th May 2019) <https://www.flawedfacedata.com/> accessed 17th November 2020.

[292] Amos Toh, 'Rules for a New Surveillance Reality' (Human Rights Watch, 18th November 2019) <https://www.hrw.org/news/2019/11/18/rules-new-surveillance-reality> accessed 17th November 2020.

[293] Roussi (n 284); Damon Wise, 'Edward Snowden on the Dangers of Mass Surveillance and Artificial General Intelligence' (Variety Technology Blog, 26th November 2019) < https://variety.com/2019/digital/festivals/idfa-edward-snowden-1203416674/> accessed 17th November 2020; Feldstein (n 277).

[294] Joseph Menn, 'Microsoft tuned down facial-recognition sales on human right concerns' (Reuters, 17th April 2019) <https://www.reuters.com/article/idUSKCN1RS2FX/#:~:text=Microsoft%20concluded%20it%20would%20lead,mostly%20white%20and%20male%20pictures> accessed 17th November 2020.

[295] Amazon, 'We are implementing a one-year moratorium on police use of Rekognition' (Amazon Policy News, 10th June 2020) <https://www.aboutamazon.com/news/policy-news-views/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition> accessed 17th November 2020.

[296] Arvind Krishna (IBM CEO), 'Letter to Congress on Racial Justice Reform' (IBM, 11th November 2020) <https://www.ibm.com/policy/facial-recognition-sunset-racial-justice-reforms/> accessed 18th November 2020; Kent Walker, 'AI for Social Good in Asia Pacific' (Google post, 13th December 2018) <https://www.blog.google/around-the-globe/google-asia/ai-social-good-asia-pacific/amp/> accessed 18th November 2020.

and have reportedly ceased or reduced the work on the technology altogether,[297] with similar positions also taken by the cities of Oakland,[298] San Francisco,[299] and Somerville[300] in the US, in concern of police and government use and the risk of bias.

The majority agree that FRT simply does not succeed in its intended purpose, is not necessary and proportionate, and is frequently misused by those who have access to it.[301] Already in the US, there has been evidence of FRT failing to identify the correct individual, leading to consequential false arrests.[302] In June 2020 in Michigan, an African-American man was arrested following a charge of larceny, after he was identified as a suspect using FRT on the CCTV footage of the crime.[303] Although this identification should only be used as a clue in police work, and not full probable cause of arrest, the photo of the individual was included in a line-up.[304] Following this, a lawyer from the American Civil Liberties Union (ACLU) of Michigan called to the Chief Investigator for the case to be dismissed without prejudice,[305] a public apology to be issued, and for FRT to stop being used as an investigatory tool by Michigan law enforcement and other agencies.[306]

In 2019, a trial which was based on this technology found evidence of racial bias,[307] which reflects the consequences where a lack of safeguarding exists to protect rights, specifically Article 14,[308] and highlights the wider effects on society that AI can

[297] David Paris, ‘Australia needs to face up to the dangers of facial recognition technology’ *The Guardian* (7th August 2020).

[298] Sarah Ravani, ‘Oakland committee ban on facial recognition surveillance’ *The San Francisco Chronicle* (25th June 2019).

[299] Kate Conger, Richard Fausset and Serge Kovaleski, ‘San Francisco Bans Facial Recognition Technology’ *The New York Times* (14th May 2019).

[300] Sarah Wu, ‘Somerville City Council passes facial recognition ban’ *The Boston Globe* (27th June 2019).

[301] Davide Castelvecci, ‘Beating Biometric Bias’ [2020] 587 *Nature* 347; Information Commissioner’s Office, *The use of live facial recognition technology by law enforcement in public places* (2019) 15-17; Pete Fussey and Daragh Murray, *Independent Report on the London Metropolitan Police Service’s Trial of Live Facial Recognition Technology* (The Human Rights, Big Data and Technology Project, 2019) 43-45.

[302] Hill (n 266).

[303] ibid.

[304] ibid.

[305] This was confirmed by the Prosecutor’s Office here: Wayne County Prosecuting Office, ‘Statement in response to New York Times Article Wrongfully Accused by an Algorithm’ (Press Release, 24th June 2020) <https://int.nyt.com/data/documenthelper/7046-facial-recognition-arrest/5a6d6d0047295fad363b/optimized/full.pdf#page=1> accessed 19th November 2020.

[306] Phil Mayor, ‘Letter to Chief Investigator’ (American Civil Liberties Union, 24th June 2020) <https://www.aclu.org/letter/aclu-michigan-complaint-re-use-facial-recognition> accessed 18th November 2020.

[307] National Institute of Standards and Technology, *Facial Recognition Vendor Test (FRVT) Part 3: Demographic Effects* (2019) 66.

[308] ECHR (n 3) Article 14.

pose if it is left unregulated. There have also been reports of misuse of FRT by law enforcement, in particular due to the lack of rules controlling what data police are permitted to submit into the technology.[309] Images which are low-quality or have filters,[310] as well as computer-generated features and artist sketches[311] can currently be used in systems, which clearly would have an effect on the output of the technology, and with evidence of bias in systems,[312] it is clear that more safeguards are needed in order to protect human rights.

2.1.4 The Case of R(Bridges) v South Wales Police

The dangers and potential devastating impact of FRT in the UK are highlighted by The Human Rights, Big Data and Technology Project, who also raise the concern of the inaccuracy of FRT, especially when in public use, in which trials by London Metropolitan Police resulted in only 19% of correct matches.[313] These concerns have also been made realistic with the world-first legal challenge of its kind, the case of *R(Bridges)*,[314] which has made a substantial impact to the consideration of future use, development and regulation of AI.

In the Court of Appeal,[315] after the previous judgment in which the police use of FRT technology was found lawful,[316] it was held that the use of the technology was unlawful and violated human rights.[317] This was due to FRT having no sufficient legal framework for its use, and therefore the interference with human rights not fulfilling the ‘prescribed by law’ requirement needed to establish a lawful infringement. The judgment also highlights the issue where too much discretion is left in the hands of law enforcement officers,[318] and raises the concern of discrimination and the lack of

[309] Garvie (n 286).

[310] Tomasz Marciniak, Agata Chmielewska, Radoslaw Weychan, Marianna Parzych and Adam Dabrowski, ‘Influence of low resolution of images on reliability of face detection and recognition’ [2015] 74 *Multimedia Tools and Applications* 4329; Mohammad Haghighat and Mohamed Abdel-Mottaleb, ‘Low Resolution Face Recognition in Surveillance Systems Using Discriminant Correlation Analysis’ [2017] *12th IEEE International Conference on Automatic Face and Gesture Recognition* 912.

[311] Nancy Abudu, ‘Letter to the Orlando Police Department’ (American Civil Liberties Union, documents on use of Amazon recognition service, January 2018) <https://www.aclunc.org/docs/20180522_ARD.pdf> accessed 3rd November 2020.

[312] Osonde Osaba and William Welser, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence* (Rand Corporation, 2017) 17; Coeckelbergh (n 205) 125; Turner (n 205) 337.

[313] Fussey and Murray (n 296) 10.

[314] *R(Bridges) v CC of South Wales Police* (n 47)*; R(Bridges) v CC of South Wales Police* (n 49).

[315] *R(Bridges) v CC of South Wales Police* (n 49).

[316] *R(Bridges) v CC of South Wales Police* (n 47).

[317] *R(Bridges) v CC of South Wales Police* (n 49).

[318] ibid [91].

adequate checks and balances to address this.[319] The judgment of this case has been declared a major victory in the fight against discriminatory and oppressive FRT,[320] and is hoped to be viewed as a deterrent to police forces rolling out oppressive technologies without sufficient consideration.[321] The judgment from the Court of Appeal was welcomed by the Surveillance Camera Commissioner, who expressed that this technology should not be in police use if it cannot be shown that it is fair and non-discriminatory.[322] The ICO also responded to the judgment, praising the clarification on police use of FRT.[323]

Worryingly, also in the Netherlands, the usage and challenges of AI technology have reached the courts.[324] This case concerned the use of the Government's risk indication system (syRi), developed over the past decade to predict the likelihood of an individual committing benefit or tax fraud, or violating related labour laws.[325] The system was deployed primarily in low-income neighbourhoods, and used past government data in order to predict its decisions. The court held that the legislation prescribing the use of SyRi did not contain sufficient privacy safeguards.[326] The judgment declared the relevant legislation invalid, as it did not fit the criteria for a 'fair balance' between its objectives; to prevent and combat fraud against the violation of privacy.[327] The lack of transparency also heavily weighted on the court's decision, especially due to the potentially discriminatory effects of the technology in question, and the lack of ability to assess this.[328]

[319] ibid [199].

[320] Liberty, 'Liberty wins ground-breaking victory against facial recognition tech' (Press Release, 11th August 2020) <https://www.libertyhumanrights.org.uk/issue/liberty-wins-ground-breaking-victory-against-facial-recognition-tech/> accessed 5th November 2020.

[321] Big Brother Watch, 'Response to Court of Appeal Judgment in Dr Bridges' Challenge' (11th August 2020) <https://bigbrotherwatch.org.uk/2020/08/big-brother-watchs-response-to-court-of-appeal-judgment-in-dr-bridges-challenge-to-live-facial-recognition/> accessed 5th November 2020.

[322] The Surveillance Camera Commissioner, 'statement on Court of Appeal judgment (R) Bridges v South Wales Police – Automated Facial Recognition' (Gov.uk, press release, 11th August 2020) <https://www.gov.uk/government/speeches/surveillance-camera-commissioners-statement-court-of-appeal-judgment-r-bridges-v-south-wales-police-automated-facial-recognition> accessed 5th November 2020.

[323] Information Commissioner's Office, *The use of live facial recognition technology in public places* (2021) 36.

[324] *Dutch Legal Committee for Human Rights v State of the Netherlands* [2020] C/09/550982 HA ZA 18-388.

[325] Philip Alston, 'Landmark ruling by Dutch court stops government attempts to spy on the poor' (United Nations Human Rights Office of the High Commissioner, UN Special Rapporteur on extreme poverty and human rights, 5th February 2020) < https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25522> accessed 10th December 2020.

[326] *Dutch Legal Committee for Human Rights v State of the Netherlands* (n 319).

[327] Alston (n 320).

[328] *Dutch Legal Committee for Human Rights v State of the Netherlands* (n 319).

Van Veen praises this decision, expressing that it sets a strong legal precedent and shows how effective human rights advocacy can be.[329] The UN Special Rapporteur on Extreme Poverty and Human Rights, Alston, also praised the landmark judgment and predicts the ruling will be a 'wake-up call' for the Governments in other countries who are experimenting with similar products to digitise themselves.[330] In reflection of these cases, it is evident that AI technology can cause serious concern to human rights, reflecting the importance of ensuring future frameworks on AI are based on adequate safeguards and protections for individuals and their fundamental rights and freedoms.

2.1.5 Discrimination and Bias

In reference to the above discussion, recommendations on reducing the risk to Article 14 are also prominent throughout the literature, not only reflected by the highlighted cases, but also due to several AI systems which have reportedly shown evidence of bias, or developing bias. In possibly one of the most public (yet amongst countless examples of AI bias), Microsoft produced an AI Chatbot named 'Tay' in 2016 as an experiment in conversational understanding to engage on Twitter (now known as X), intending to allow the system to develop and learn through engagement with the public online.[331] Within 16 hours, Tay began to tweet racist, sexist and rude remarks, leaving Microsoft little choice but to suspend the account.[332] This bias in the system was developed due to the large amounts of negative tweets being received, in which the words used were consequently learnt and reproduced by Tay.[333]

Several independent studies have also highlighted the risk of bias in systems, including the Massachusetts Institute of Technology (MIT) study in 2018 in which FRT incorrectly identified up to 35% of darker-skinned women, and it was found that

[329] Christiaan Van Veen, 'Landmark judgment from the Netherlands on digital welfare states and human rights' (Open Global Rights Blog, 19th March 2020) <https://www.openglobalrights.org/landmark-judgment-from-netherlands-on-digital-welfare-states/> accessed 10th December 2020.
[330] Alston (n 320).
[331] Peter Lee, 'Learning from Tay's introduction' (Microsoft Blog, 25th March 2016) <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/> accessed 8th November 2020.
[332] Oscar Schwartz, 'In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation' (IEEE Spectrum Blog, 25th November 2019, updated 4th January 2024) <https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation> accessed 8th November 2020.
[333] ibid.

police use of FRT disproportionately affects African-Americans.[334] It has been found consistently that the sensors recognised lighter-skin tones better than darker-skin tones, due to being developed and tested on lighter-skin more often.[335] These examples reflect one way where bias can exist within AI systems, in which already biased data is received by or input into the system, making AI replicate that same bias. This risk emphasises the need for additional safeguards regarding the data inputted into devices, and the need to mitigate and monitor the risk of bias in AI systems. Legislative rules that enforce diverse data to be used in the training dataset of AI systems, and assurances that these are complied with are a necessity to ensure the best protection of human rights.

The second way bias can develop in systems is through the technology developing a bias itself, through categorising or prioritising unrelated variables of data, highlighting the need for stricter technical rules and legislation on the development of AI to reduce the risk of this occurring. For example, a hiring algorithm by Amazon was found to favour applicants who had words such as ‘executed’ or ‘captured’ on their resume, and these words were more commonly found on men’s resumes making the system favour males.[336] Due to this, it is imperative that there are legislative rules that monitor and set limitations on the use and deployment of AI, as well as additional safeguards to protect human rights.

Amnesty International and Access Now are admired in relation to this aspect due to their involvement in and the subsequent release of the Toronto Declaration; a landmark report on protecting the right to equality and non-discrimination in ML systems.[337] Saucedo expresses that the correct regulations and safeguards will be a catalyst for AI innovation, and suggests that a way to address bias is to implement

[334] Larry Hardesty, ‘Study finds gender and skin-type bias in commercial artificial-intelligence systems’ (MIT News, 11th February 2018) <https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212> accessed 10th November 2020.

[335] Anthony Cuthbertson, ‘Self-Driving Cars more likely to drive into black people, study claims’ *The Independent* (6th March 2019).

[336] Jeffrey Dastin, ‘Amazon scraps secret AI recruiting tool that showed bias against women’ (Reuters Blog, 11th October 2018) <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> accessed 4th November 2020.

[337] Amnesty International and Access Now, *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems* (2018).

boundaries on the results AI can produce.[338] Shown in practice, using this method for predicting a college class exam results, the user can impose a limit on the amount an algorithm can overestimate and underestimate the scores for males and females on average, thus reducing the risk of a bias developing where one sex is favoured over the other.[339] However, without mandatory requirements for suggestions like this to be included when developing an algorithm, it can be assumed that the majority of developers would not impose, in their view, such a restriction on their products.

Clear policies and good practices are needed for the data inputted into AI systems, especially in scenarios where companies do not have enough data internally to recognise a bias or have reason to believe it may contain bias.[340] The author of Technology Ethics, Hare, calls for AI ethics to be codified in a realistic and enforceable way,[341] similar to Manyika, Silberg and Presten who highlight the need for accelerated progress in addressing bias in AI, commenting that there is no 'quick fix' to making systems fair.[342] The understanding of 'fairness' poses the biggest challenge is relation to this aspect, and although developers may ensure that the systems satisfy their definition of fairness, it may not be sufficient for others.[343] Verma and Rubin acknowledge over twenty differing definitions of fairness,[344] highlighting an area which may benefit from clarification within legislation, to provide a standard and consistent consensus of how to achieve fairness, or minimum requirements to ensure the reduction of unfairness in AI systems. For developers to

[338] Alejandro Saucedo, 'The top three risks posed by AI, and how to safeguard against them' (IR Pro Portal Blog, 14th August 2020) <https://www.seldon.io/itproportal-the-top-three-risks-posed-by-ai-and-how-to-safeguard-against-them> accessed 5th November 2020.

[339] Philip Thomas, Bruno Castro da Silva, Andrew Barto, Stephen Giguere, Yuriy Brun and Emma Brunskill, 'Preventing undesirable behaviour in intelligent machines' [2019] 366(6468) *Science* 999.

[340] Reuben Binns and Valeria Gallo, 'Human bias and discrimination in AI systems' (Wired Gov Blog, 26th June 2019) <https://www.wired-gov.net/wg/news.nsf/articles/Human+bias+and+discrimination+in+AI+systems+26062019152000?open> accessed 5th November 2020.

[341] Stephanie Hare, 'It's time for AI ethics to grow up' (Wired Blog, 8th January 2020) <https://www.wired.co.uk/article/ai-ethics-law> accessed 4th November 2020.

[342] James Manyika, Jake Silberg and Brittany Presten, 'What Do We Do About the Biases in AI?' (Harvard Business Review, 25th October 2019) <https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai> accessed 5th November 2020.

[343] Jon Kleinberg, Senhil Mullainathan and Manish Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' (Cornell University, 2016) <https://arxiv.org/abs/1609.05807v2> accessed 5th November 2020; Alexandra Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments' [2017] 5(2) *Big Data* 153.

[344] Sahil Verma and Julia Rubin, 'Fairness Definitions Explained' (International Workshop on Software Fairness, Sweden, Fairware, 2018) <https://fairware.cs.umass.edu/papers/Verma.pdf> accessed 12th August 2022.

be able to identify bias within their AI systems, they arguably need to be able to understand the decisions made by AI technology to enable them to address it.[345]

2.1.6 Transparency, Explainability and Interpretability

In relation to the above discussion, the literature considers the concepts of transparency, explainability and interpretability in vast amounts to provide solutions to mitigate bias and unfairness.[346] Transparency in AI is understood broadly as the ability to determine how and why an algorithm arrived at its decision.[347] The concept of transparency is widespread throughout various disciplines, and is seen as a clear standard requirement for the future of AI.[348] Transparency is included as one of the seven key requirements for 'trustworthy AI', proposed in the EC's HLEG on AI's report,[349] and is one of the most common elements included in recommendations towards regulation for AI.[350] Some level of transparency needs to be required in order to ensure a future of ethical AI, to recognise incorrect decisions as well as identifying bias,[351] in addition to aiding the placing of responsibility.[352]

The fundamental tensions between different objectives need to be acknowledged in relation to demanding 'too much' transparency, including the conflict with privacy protections and corporate competition, as well as increasing the risk of misuse by users, or those 'gaming' AI.[353] In a study conducted in 2018, transparent AI models resulted in achieving the opposite effect, with mistakes and errors being harder to spot and therefore fix, highlighting the possibility of transparency leading to an information overload.[354] Although transparency is important for the future of AI, the concept alone will make little effect. AI does not think or work in the same way as

[345] Saucedo (n 333).
[346] ibid.
[347] Larsson and Heintz (n 182).
[348] ibid; Felzmann, Fosch-Villaronga, Lutz and Tamo-Larrieux (n 184); Heike Felzmann, Eduard Fosch-Villaronga, Christopher Lutz and Aurelia Tamò-Larrieux, 'Towards Transparency by Design for Artificial Intelligence' [2020] 26(2) *Science and Engineering Ethics* 3333*.*
[349] High-Level Expert Group on Artificial Intelligence (n 100) 14.
[350] Leilani Gilpin, David Bau, Ben Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal, 'Explaining Explanations: An Overview of Interpretability of Machine Learning' (IEEE 5th International Conference on Data Science and Advanced Analytics, 2018) <https://arxiv.org/abs/1806.00069> accessed 20th July 2021.
[351] ibid.
[352] Deloitte, *Transparency and Responsibility: A call for explainable AI* (2019) 6.
[353] Larsson and Heintz (n 182).
[354] Forough Poursabzi-Sangdeh, Daniel Goldstein, Jake Hofman, Jennifer Vaughan and Hanna Wallach, 'Manipulating and measuring model interpretability' (Cornell University, 8th November 2020) <https://arxiv.org/abs/1802.07810> accessed 6th November 2020.

humans, and therefore complete transparency would only amount to confusion, and may not be understandable at all, failing to achieve what transparency originally sets out to do.[355] The importance of interpretability is underpinned in relation to this aspect,[356] to enable the average user to understand the workings behind AI.

In connection to this, the literature discusses explainability and interpretability as neighbouring concepts to transparency.[357] Policy debates across the world are increasingly calling for some form of AI explainability, to increase the ethical standards with systems.[358] In the Royal Society's Policy Briefing,[359] the reasons as to why explainability and interpretability are deemed necessary and desirable are noted, including to increase public trust and confidence in AI.[360] Other reasons include safeguarding against bias, enforcing legal rights, and assessing the workings of a system and its vulnerabilities, to meet the expectations of society.[361] It is accepted throughout the literature that in scenarios where there are no significant consequences from AI decisions, much less concern in relation to explainability exists.[362]

On the opposite end of the spectrum, it is well debated that for higher risk AI systems, although explainability may be necessary, it may not be effective for addressing accountability matters sufficiently, reflecting that transparency, explainability and interpretability are only the initial steps in creating trustworthy AI.[363] The level of explainability and whether a right to an explanation exists are also debated in the literature.[364] The Royal Society note that different users require different forms of explanation in different contexts in order to understand them,[365] in addition to the issue of unreliable explanations, and that explainability alone cannot answer questions about accountability.[366]

[355] Gilpin, Yuan, Bajwa, Specter and Kagal (n 345).
[356] ibid.
[357] ibid.
[358] The Royal Society, *Explainable AI: The Basics* (Policy Briefing, November 2019) 21-23.
[359] ibid.
[360] ibid 9; Jake Silberg and James Manyika, Notes from the AI frontier: *Tackling bias in AI (and in humans)* (McKinsey Global Institute, 2019) 6; Saucedo (n 333).
[361] Silberg and Manyika (n 255) 2-6.
[362] The Royal Society (n 353) 24.
[363] Chris Oxborough and Euan Cameron, 'Explainable AI' (PWC, 2018) <https://www.pwc.co.uk/services/risk-assurance/insights/explainable-ai.html> accessed 3rd November 2020.
[364] Goodman and Flaxman (n 195); Wachter, Mittelstadt, and Floridi (n 195); Edwards and Veale (n 179).
[365] The Royal Society (n 353) 19.
[366] ibid 23.

A right to explanation is arguably included within the GDPR,[367] with the intention of addressing the potential issues stemming from the use of AI systems. The extent of this right is heavily debated throughout the literature,[368] in which the major criticism is concentrated in two areas, firstly that it is not included within the binding parts of the legislation after being removed during the legislative process.[369] Secondly, concerns are raised due to the limitations of the right, as it only applies to decisions 'solely' based on automated processing, and have resulting legal or similarly significant consequential effects, making this right unlikely to apply in several circumstances.[370] A view that is less prominent in the literature is that a 'right to explanation' is not needed, and could potentially stifle innovation.[371]

2.1.7 Future Regulation

To address the rising concerns between AI and human rights, new hard regulation specifically for this technology is needed, to ensure the best protection to society. Privacy International and Article 19 express that the development, use and research of AI should be at least *"subject to the minimum requirement of respecting, promoting and protecting international human rights standards*",[372] and advocate for "*AI human right critical systems*". [373] The UN also express that any effort to develop policy or regulation in the field of AI needs to ensure consideration of human rights concerns, highlighting the UK's detailed report on AI in 2018 which fails to consider human rights once.[374]

With AI being used in a variety of industries, either in public use, or to aid private businesses, services and decision-making, the potential lasting effects this technology will have on every human right is understated. If AI technology continues to develop and transform industries and society, it is inevitable that new issues and

[367] GDPR (n 1) Recital 71.
[368] Goodman and Flaxman (n 195); Wachter, Mittelstadt, and Floridi (n 195).
[369] Wachter, Mittelstadt, and Floridi (n 195).
[370] ibid; Edwards and Veale (n 179).
[371] Nick Wallace, 'EU's Right to Explanation: A Harmful Restriction on Artificial Intelligence' (TechZone Blog, 25th January 2017) <https://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm> accessed 5th November 2020.
[372] Privacy International and Article 19 (n 247) 28.
[373] ibid 28.
[374] United Nations (n 82) 15; House of Lords, *Order of Business* (Volume 794, No. 208, Select Committee Report on Artificial Intelligence, 19th November 2018).

challenges will arise which are not yet known, and it is imperative that the scope of future regulation has the capabilities of dealing with this. This theory is currently underrepresented in the literature, where the focus is centred on tackling the current issues and challenges of AI. One of the ways this could be made possible is allocation in legislation for future funding for research, and whilst Access Now include this in their report,[375] it is not considered by many others. In the next part of this review, the literature surrounding the liability issues of AI are explored, to highlight that the current liability principles are not sufficient to address the new complexities that AI technology brings.

2.1.8 AI and Liability

In addition to the human right and ethical concerns, another regulatory challenge posed by AI is the issue of liability. As AI systems develop, integrate further into society and become more autonomous, it needs to be addressed how responsibility should be allocated. This section of the literature review seeks to identify the relationship between AI and placing liability. AI brings new complexities to standard liability frameworks mainly due to the issue of foreseeability, which is discussed extensively throughout the literature.[376] Whilst traditional liability principles have served well in regulating human actions, the unpredictable and ML nature of AI introduces new challenges that the current liability frameworks do not sufficiently address. Although each liability option has its own challenges, various ideas are proposed in the literature, including the tort of negligence,[377] the PLD,[378] vicarious

[375] Access Now (n 275) 32.

[376] Kowert (n 71); Carlos Gaviria, ‘The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps Generated by AI in the US’ (Pardee Rand, 2020) <https://www.rand.org/pubs/rgs_dissertations/RGSDA319-1.html> accessed 10th August 2021; Turner (n 205) 61; European Commission (n 67) 44.

[377] Hubbard (n 221); Kingston (n 278); Reed (n 181).

[378] Cauffman (n 209); Yaniv Benhamou and Justine Ferland, ‘Artificial Intelligence and Damages: Assessing Liability and Calculating the Damages’ (Leading Legal Disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law, 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3535387> accessed 20th June 2021.

liability,[379] a personhood status for AI,[380] use of insurance schemes, [381] and use of a no-fault scheme.[382]

In addition to these options, there is evidence of companies offering the option of self-regulation.[383] The car manufacturer Volvo expressed in 2015 that they will accept full liability where an incident occurs during one of its cars being at fault and in autonomous mode, being one of the first car makers to make such a promise.[384] Expressing this clearly increases public trust in the company, and although only announced through a press release, if the courts were to follow the precedent of *Carlill v Carbolic Smoke Company*,[385] this promise may be held to be binding. However, there is still a high likelihood of systems that are not subject to their own self-regulation going wrong, and therefore self-regulation will not solve the full liability issue alone, but does pose an option for those companies who seek to stand out.

It is also important to note the relevance of human rights when discussing tortious liability, whereby case-law (predominantly in areas such as breach of confidence and misuse of private information) has shown a clear interplay between the two. For example, in one of the most well-known cases, the safeguarding of Article 8 was seen as the core priority within the case of *Campbell,*[386] reflecting the broader role that claims can play in ensuring accountability and redress for harm, particularly

[379] Paulius Čerka, Jurgita Grigienė and Gintarė Sirbikytė, 'Liability for damages caused by artificial intelligence' [2015] 31(3) *Computer Law and Security Review* 376; Lynn Richmond, 'Artificial Intelligence: Who's to blame?' (Tech Law for Everyone, 8th August 2018) <https://www.scl.org/articles/10277-artificial-intelligence-who-s-to-blame> accessed 12th November 2020.

[380] Janosch Delcker, 'Europe divided over robot 'personhood' (Politico Blog, 11th April 2018) <https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/> accessed 20th March 2019; Robotics Openletter, 'Open Letter to the European Commission, Artificial Intelligence and Robotics' (2018) <https://robotics-openletter.eu/#:~:text=We%2C%20Artificial%20Intelligence%20and%20Robotics,Union%20citizens%20while%20fostering%20innovation.> accessed 10th June 2021; Jiahong Chen and Paul Burgess, 'The boundaries of legal personhood: how spontaneous intelligence can problematize differences between humans, artificial intelligence, companies and animals' [2019] 27(1) *Artificial Intelligence and Law* 73; European Parliament, *Report with recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103(INL)) 59F.

[381] Turner (n 205) 113; European Commission (n 67) 63; Curtis Karnow, 'Liability for Distributed Artificial Intelligences' [1996] 11(1) *Berkley Technology Law Journal* 147.

[382] Turner (n 205) 103.

[383] Volvo, 'US urged to establish nationwide Federal guidelines for autonomous driving' (Press Release, 7th October 2015) <https://www.media.volvocars.com/global/en-gb/media/pressreleases/167975/us-urged-to-establish-nationwide-federal-guidelines-for-autonomous-driving> 11th November 2020; John McIlroy, 'Toyota boss says autonomous cars will accept crash liability' (AutoExpress, 22nd October 2019) < https://www.autoexpress.co.uk/toyota/108205/toyota-boss-says-autonomous-cars-will-accept-crash-liability> accessed 13th November 2020.

[384] Volvo (n 378).

[385] *Carlill v Carbolic Smoke Ball Company* [1892] EWCA Civ 1.

[386] *Campbell v Mirror Group Newspapers Ltd* [2004] UKHL 22 [2004] 2 WLR 1232.

where that harm engages fundamental human rights. Continuing, this section seeks to explore the guidance issued by the EC on AI and liability,[387] in addition to highlighting the connecting issues in relation to the most common suggestions of liability for AI.

Liability for AI

In one of the most encouraging sources, the HLEG on Liability and New Technologies, on behalf of the EC, released a detailed report on liability for AI and other emerging digital technologies.[388] This expert group concluded that the liability regimes in force ensure at least basic protection to parties seeking redress, however, acknowledge that specific characteristics, including complexity, self-learning capabilities and predictability may make it more difficult for victims to claim where it may be justified.[389] The report lists a number of findings in which current liability regimes should be redesigned, highlighting that it is difficult to allocate fault-based liability rules due to the lack of well-established models of a 'proper function' standard, in addition to the likelihood of unforeseeable actions by AI systems.[390]

This view was consistent in the EU's White Paper for AI, which stated that there is a need to examine whether current liability frameworks are able to address the risks and challenges brought by AI, and whether they can be effectively enforced through adaptions to legislation, or if new legislation is needed altogether.[391] The EU's report on liability for AI[392] considers many options which are examined in this section of the literature review, however arguably un-surprisingly, it does not consider the final option of a widespread no-fault scheme, except for in times where other claims cannot be satisfied.[393]

Negligence and AI

One of the major themes that exists in the literature when exploring AI liability is the tort of negligence. The modern law of negligence was established in *Donoghue v*

[387] European Commission (n 67).
[388] ibid.
[389] ibid 3.
[390] ibid 23.
[391] European Commission (n 112) 10.
[392] European Commission (n 67).
[393] ibid 9.

Stevenson,[394] in which clarification was given in regard to who amounted to a 'neighbour', in that "*reasonable care must be taken to avoid acts or omissions which can reasonably foresee be likely to cause injury to a neighbour*".[395] A 'neighbour' was clarified to be "*any persons who are closely and directly affected by the act*",[396] and ought reasonable to be considered as being affected.[397] The use of the term 'reasonable' highlights the issue of foreseeability, which brings complexities when dealing with AI.

In order to prove a successful negligence claim, there must be evidence that the defendant owed a duty of care, that the defendant was in breach of that duty, that the breach caused the damage, and that the damage was not too remote.[398] The legal test for establishing a duty of care differs depending on the type of loss,[399] in which foreseeability is a common theme throughout. When considering AI, several questions can be posed, firstly, whether a duty of care should exist when dealing with AI, and whether this should lie with the developer or the deployer of the system. The second question; what amounts to a breach of this duty, and how can it be measured? Thirdly, how to establish proof that the breach has a link to the damage caused, and finally, whether that damage was reasonably foreseeable, and whether foreseeability is an adequate standard?[400] The advantages to using negligence is that there is no limit on who can make a claim, and that a higher standard of duty of care is placed on those with a higher risk of damage, making the rules work proportionately.[401]

The central concept of negligence is whether the defendant acted reasonably, or the same as an average, reasonable person would in that given situation, historically compared to how "*a man on a Clapham omnibus"* would act.[402] This test is problematic when used with humans using AI, or even AI itself,[403] giving rise to the

[394] *Donoghue v Stevenson* (n 59).
[395] ibid.
[396] ibid.
[397] *Donoghue v Stevenson* (n 59) [580].
[398] ibid (n 59) [619]; *Caparo Industries plc v Dickman* (n 59) [632]; *The Wagon Mound no 1* (n 59).
[399] For Duty of Care: *Donoghue v Stevenson* (n 59) [619]; *Caparo Industries plc v Dickman* (n 59) [632]; For psychiatric injury: *Page v Smith* [1996] 1 AC 155; For causation: *Barnett v Chelsea & Kensington Hospital* [1969] 1 QB 428; For remoteness of damage: *The Wagon Mound no 1* (n 59).
[400] Turner (n 205) 86.
[401] ibid 88.
[402] *Mcquire v Western Morning News* [1903] 2 KB 100 [109].
[403] Turner (n 205) 88.

question of how and who should set standards for AI behaviour. Several AI systems include aspects of ML, which "*departs from software coding in the conventional sense and begins to look more like coaching that it does programming*".[404] This means that as the AI technology interacts with society, it calculates its most successful results of their actions for use in future decisions, evolving through time.[405] This ability increases the likelihood of AI acting unpredictably, and gives rise to the risk of a lack of control of systems once they are sold.[406] The more unpredictable, or unforeseeable these actions are, the more complex it is to assign liability, as developers can depend on the foreseeability clause needed for a negligence claim, and therefore avoid responsibility.[407]

Abbot provides a solution to this issue, proposing that if a manufacturer can show that an autonomous machine is safer in comparison to a 'reasonable person', but subsequently causes harm or damage through negligent actions, the supplier should be liable under negligence, instead of strict liability.[408] Abbott's proposed test focuses on AI's "*act instead of its design, and in a sense, it would treat a computer tortfeasor as a person rather than a product*", determining negligence on the standard of a 'reasonable computer'.[409] This raises additional questions as to assigning such a standard, by which Abbot addresses should be considered on the industry custom, average, or safest technology.[410] This solution, although somewhat encouraging, would need to be expanded in order to apply to differing applications of AI technology, as one system's 'average' may differ substantially to another. The terms 'reasonable' and 'foreseeable' are the major concerns in relation to assigning AI with negligence claims, and due to these being so complex, several other options are also discussed in the literature.

[404] Kowert (n 71).
[405] Jason Tanz, 'Soon We Won't Program Computers. We'll Train Them Like Dogs' (Wired Blog, 17th May 2016) <https://www.wired.com/2016/05/the-end-of-code/> accessed 11th November 2020.
[406] Jonathan Tapson, 'Google's Go Victory Shows AI Thinking Can Be Unpredictable, and That's a Concern' (The Conversation, 17th March 2016) < https://theconversation.com/googles-go-victory-shows-ai-thinking-can-be-unpredictable-and-thats-a-concern-56209> accessed 11th. November 2020.
[407] Scherer (n 70).
[408] Ryan Abbott, 'The Reasonable Computer: Disrupting the Paradigm of Tort Liability' [2017] 86(1) *The George Washington Law Review* 101.
[409] ibid.
[410] ibid.

AI and Product Liability

As noted, traditional regimes such as negligence often rely on concepts such as foreseeability, which can be problematic when applied to AI systems. These complexities encourage the consideration of other, alternative frameworks which may better suit AI's specific characteristics. Often seen as controversial,[411] strict liability is a form of liability which does not consider fault, and instead, allocates liability depending on the act in question. The benefits of a strict liability regime include ensuring victims are properly compensated, and provides legal certainty, which in turn encourages precaution by deployers and forces developers to increase the safety and control of their systems. The EU has commented that strict liability is an appropriate response to address the risks of AI in public environments, however questions whether the defences and statutory exceptions from strict liability would need to be re-considered.[412]

One form of strict liability is product liability for defective products, set out in the PLD,[413] in which liability is established on whether the injured person can prove the defect, the damage caused, and that a causal relationship existed between them. Article 6 of the PLD describes a product as defective when "*it does not provide the safety which a person is entitled to expect, taking into account circumstances including the presentation of the product, the use to which it could reasonably be expected that the product would be put, and the time when the product was first circulated*".[414] This test for defectiveness is viewed in the literature as open-ended,[415] especially in comparison to the more structured approach in the US in which the defect in question must be related to either the design, instruction and/or warnings, or manufacturing.[416]

[411] Turner (n 205) 91; Rod Freeman, Claire Temple, Tracey Bischofberger, Sarah-Jane Dobson and Carol Roberts, 'Product liability and safety in the EU: overview' (Practical Law, 1st August 2020) <https://uk.practicallaw.thomsonreuters.com/w-013-0379?transitionType=Default&contextData=(sc.Default)&firstPage=true> accessed 13th November 2020.
[412] European Commission (n 67) 39.
[413] Product Liability Directive (n 66) 7.8.
[414] ibid 7.8 and Article 6.
[415] Turner (n 205) 93.
[416] Lord Griffiths, Peter De Val and R.J Dormer, 'Developments in English Products Liability Law: A comparison with the American system' [1988] 62 *Tulane Law Review* 354; *A and Others v National Blood Authority and another* [2001] 3 All ER 289.

Article 7 of the PLD details the circumstances where the producer may not be liable, including scenarios where the defect did not exist at the time the product was put into circulation, or that knowledge at the time the product was put into circulation did not exist to identify the defect.[417] This Article is seen as problematic in the literature when considering product liability for AI, as the PLD is based on the assumption that products are static,[418] rather than having self-learning capabilities. This poses the question of whether an amendment to the PLD would be more sufficient in covering the complexities AI technology brings in comparison to negligence, due to the foreseeability issue not existing.

Another area of concern includes whether AI is to be considered a product or a service, and therefore if it would be able to fall into the PLD at all. Article 2 of the PLD defines a ‘product’ as ‘all movables’, which in relation to AI would include robotic products, but not cloud-based systems,[419] and therefore would not be sufficient. Article 2 does explicitly include electricity as a product,[420] so it could be assumed that this could also include the more abstract AI systems, however further clarification and amendment would be needed to ensure this. To address these concerns, the EU propose amendments to the current PLD to add clarity and refine the rules for redress;[421] these amendments are discussed in further detail in the review of the recent regulatory landscape within Chapter Five. Next in this review, the literature surrounding AI and vicarious liability will be explored.

AI and Vicarious Liability

Vicarious liability is a form of strict, secondary liability that exists when one party (the principal) has responsibility over another party (the agent), with examples of an employer and employee, a parent and child, or a teacher and student.[422] Vicarious liability differs from strict liability in that not every action made by the agent will render the principle liable. Firstly, a recognised relationship needs to be established

[417] Product Liability Directive (n 66) 7.8 and Article 7.

[418] Turner (n 205) 98.

[419] ibid 95.

[420] Product Liability Directive (n 66) 7.8 and Article 2.

[421] European Parliament, *Artificial Intelligence Liability Directive* (Briefing, EU Regulation in Progress, February 2023).

[422] *Yewens v Noakes* (1881) 6 QBD 530; *Lister v Hesley Hall Ltd* [2001] UKHL 22 [2001] 5 WLUK 105; *Honeywill and Stein Ltd v Larkin Brothers Ltd* [1934] 1 KB 191; *Mohamud v WM Morrison Supermarkets Plc* [2016] UKSC 11 [2016] 3 WLUK 90; *Cox v Ministry of Justice* [2016] UKSC 10 [2016] 3 WLUK 91.

between the principal and the agent, and the wrongful act must take place in the scope of that relationship.[423] Academics express that a relationship could be established between an AI system, and the person of who it acts for, is in disposal of, or supervises over of the system.[424] Put more simply, the developer or deployer of the AI system as the principle, and the AI system as the agent. Turner expresses that vicarious liability would strike a balance between respecting the independent agency of AI and holding a legal person responsible.[425] However, a considerable gap could arise, especially in reflection of ML unpredictability.

Firstly, acknowledged by the EC, vicarious liability requires the agent to commit wrongful conduct, and in relation to AI, it can be questioned as to how a benchmark for such conduct should be set.[426] Although not coming to a final conclusion, the EC highlights Abbott's suggestion as 'most convincing'.[427] Abbott suggests assessing conduct primarily on the same benchmark as humans, and if AI systems were able to outperform humans in terms of preventing harm, it should be compared to other similar products on the market.[428] The latter part of this suggestion may be more difficult in practice, and would require additional safeguards to ensure enough systems are available for a fair and accurate comparison to be made, as well as accounting for situations in which widespread major corporations who deploy AI commit wrongdoing, which in turn may reduce the standard of similar 'products on the market'. Also, acknowledgement needs to be made in situations where AI acts to achieve a positive end goal, but commits a wrongful act whilst doing so, in light of the fact that AI does not possess 'human qualities' such as common sense, and therefore the argument can be made that there should not be a comparison to humans at all.

Also, vicarious liability only applies to those actions which take place in the scope of the prescribed relationship. Depending on the level of autonomy possessed by the AI, it is likely that the future will lead to an increase of cases whereby the action cannot be proven to be in the scope of the said relationship, leading to further

[423] Turner (n 205) 101.
[424] Čerka, Grigienė and Sirbikytė (n 374).
[425] Turner (n 205) 101.
[426] European Commission (n 67) 48.
[427] ibid 48.
[428] Abbot (n 402).

complexities and confusion. This can be compared to the example of the relationship between a parent and child, in which at a certain point, the relationship (for liability purposes) is removed, due to the child developing their own autonomy and therefore being responsible for themselves. Using this example, it could be considered that AI systems, especially in the future may reach a level of autonomy in which their principal can no longer be responsible for it, which leads and connects to the personhood debate for AI.[429]

Personhood for AI

The concept of a personhood dates back to the 13th century when the status was granted to monasteries. In the majority of countries around the world, this model also applies to companies, meaning that corporations have some of the legal rights and responsibilities of a human being, and are considered a separate entity.[430] Legal personhoods have been granted around the world, including to specific ships, aeroplanes, animals and even a river in New Zealand, which supplies the basis of the idea of granting AI such a status.[431] A similar concept being applied to AI is argued to be less about giving robots rights, but rather about holding systems at fault if things were to go wrong.[432]

The ethical argument of granting a personhood to AI has been debated in various fields. Chen and Burgess are in favour of AI in an institutional sense having a ‘personhood’, stating that it would ease the issue of dealing with liability, allowing machines to have a separate entity with separate rules and responsibilities, such as the current stance of companies in the UK.[433] This status has been applied several times sufficiently throughout the courts, making it encouraging to believe that treating AI as a separate non-human entity in the same way could lead to similar success. This principle in company law was initially laid out in *Salomon v Salomon*, and clearly explains that the company is separate from its directors, members, and

[429] Delcker (n 375); Robotics Openletter (n 375).
[430] *Salomon v Salomon & Company Ltd* [1897] AC 2; In the US: *Santa Clara County v Southern Pacific Railroad.* 118 US 394 (1886); *Citizens United v Federal Election Commission.* 558 US 310 (2010).
[431] Eleanor Ainge Roy, ‘New Zealand river granted same legal rights as human being’ *The Guardian* (London, 16th March 2017).
[432] Delcker (n 375).
[433] Chen and Burgess (n 375); *Salomon v Salomon* (n 424).

shareholders.[434] This decision allowed the owners to be separated from the responsibilities of the company and relieve them of liability under the company's name. In respect of this, Delcker comments that there are many economic, political, and legal reasons to grant AI machines a personhood status in this same way.[435]

In 2017, Sophia the Robot, an intelligent human-like robot created by Hanson Robotics, was granted Citizenship in Saudi Arabia. [436] This decision was met by surprise worldwide, and in some cases, disbelief.[437] The idea of granting a personhood status renews questions about issues of responsibility and liability alike, as well as creating a separate ethical discussion within society.[438] In early 2017, a report from the European Parliament (EP) suggested the idea that AI machines could be granted 'electronic personalities.'[439] The reasoning behind this was to allow these machines to be insured individually and be held liable for damages if something were to go wrong,[440] however, the EC's 2018 document contained no reference to any electronic personhood,[441] which reflects a change in stance within the EU institutions.

Some academics were pleasantly surprised that electronic personhoods were being considered, with the belief that intelligent machines have the potential to merit such a personhood.[442] Several manufacturers are included in those who support granting AI a personhood status,[443] believing the granting of a personhood status equates to

[434] *Salomon v Salomon* (n 424).
[435] Delcker (n 375).
[436] Hussein Abbass, 'An AI professor explains: three concerns about granting citizenship to robot Sophia' (The Conversation, 30 October 2017) <https://theconversation.com/an-ai-professor-explains-three-concerns-about-granting-citizenship-to-robot-sophia-86479> accessed 13th February 2019.
[437] British Council, 'Should robots be citizens?' (British Council website) <https://www.britishcouncil.org/anyone-anywhere/explore/digital-identities/robots-citizens> accessed 14th November 2020; Ugo Pagallo, 'Vital, Sophia, and Co. – The Quest for the Legal Personhood of Robots' [2018] 9(9) *Information (Switzerland)* 230.
[438] European Parliament (n 375).
[439] ibid, 59F.
[440] ibid, 59F.
[441] Ugo Pagallo, 'Apples, Orange, Robots: four misunderstandings in today's debate on the legal status of AI systems [2018] 376(2133) *Philosophical Transactions of the Royal Society A* 1.
[442] Tyler Jaynes, 'Legal personhood for artificial intelligence: citizenship as the exception to the rule' [2020] 35 *AI & Society* 343; Simon Chesterman, 'Artificial Intelligence and the problem of autonomy' [2020] 1 *Notre Dame Journal of Emerging Technologies* 210; Ying Hu, 'Robot Criminals' [2019] 52 *Michigan Journal of Law Reform* 487.
[443] Peter Asaro, 'Robots and Responsibility from a Legal Perspective' (IEEE Conference on robotics and automation, Robo-ethics, Rome, 2007) <https://www.peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf> accessed 20th January 2021.

common sense, and are hopeful that such status for AI would result alike to the status of companies, which has been successful worldwide.[444]

Opposing this, the European Economic and Social Committee and UNESCO's COMEST along with over 150 academics signed an open letter, expressing that they are completely against the idea of granting a personhood to AI, due to the consequential effects relating to AI being viewed in a way they believe it should not be.[445] The EU have since made it clear that for the purposes of liability, it is not necessary for autonomous systems to have a legal personality, or any similar status.[446] UNESCO's COMEST agree with this stance in their report, where it is stated that granting AI a personhood status would be highly unreasonable, unless they possess additional qualities typically associated with humans, such as freedom of will, self-consciousness or moral agency.[447] This could indeed be a consideration in the future, where academics predict Artificial General Intelligence (AGI) will be achieved.[448]

Insurance Schemes and AI

As mentioned in the above discussion, the EP once suggested the idea of AI systems being granted an 'electronic personality', to ease liability through the use of insurance schemes.[449] The EU comments that compulsory liability "*should not be introduced without careful analysis of whether it is really needed*".[450] In the literature, the idea of using the concept of insurance highlights another option when considering not only the liability, but also the regulation of AI.[451] Insurance schemes cater to issues of uncertainty,[452] with the most obvious example being the UK system

[444] Delcker (n 375).

[445] UNESCO's COMEST, *Report of COMEST on Robotic Ethics* (Paris, September 2017); Robotics Openletter (n 375); Robonarratives, 'My response to the EU's intentions to grant robots 'Electronic Personhood' and the Open Letter to the European Commission' (By a scientist that signed the Open Letter, 19th April 2018) <https://robonarratives.wordpress.com/2018/04/19/my-response-to-eus-intentions-to-grant-robots-electronic-personhood-and-the-open-letter-to-the-european-commission/> accessed 29th March 2019.

[446] European Commission (n 67) 38.

[447] UNESCO's COMEST (439).

[448] Phil Torres, 'The possibility and risks of artificial general intelligence' [2019] 75(3) *Bulletin of the Atomic Scientists* 105

[449] European Parliament (n 375) 59F.

[450] European Commission (n 67) 63.

[451] Turner (n 205) 113; Karnow (n 376).

[452] Turner (n 205) 115.

on the use of vehicles, in which individual insurance is mandatory for driving a vehicle.[453]

US Judge Karnow expresses that insurance schemes would be the best way of dealing with AI liability, posing the suggestion that developers seeking coverage be able to submit their system to a certification procedure, and if successful, would be quoted a rate depending on the probable risk of the system.[454] He adds that the risk would be assessed along a spectrum of automation, whereby the higher the intelligence, the higher the risk, and therefore the higher rate.[455] Arguably the 'higher intelligence' part needs to be reconsidered here, following the theory that the more intelligence AI possesses, the more intelligent it will be in making safer decisions. This therefore would make it less likely for such technology to act out of turn, and therefore may be cause for a reduction in rate.

It is suggested that insurance schemes would ease and reduce the concern of situations whereby an AI system acts in an unpredictable manner,[456] however, it would not change or solve the underlying legal issue of responsibility. If an AI system were to be un-insured or was used (or acted) in a way which was not covered under the terms in the insurance scheme, there needs to be rules as to who is liable, to ensure victims are sufficiently compensated. Continuing, no-fault compensation schemes are also explored.

No Fault Liability Schemes

Less considered in the literature in comparison, but still subject to some discussion, is the idea of no-fault liability schemes.[457] Having a no-fault liability scheme would remove the legal questions of responsibility and liability, with the use of a compensation scheme that pays out irrelevant of fault, somewhat a simpler suggestion to those explored above. Although rarely used, an encouraging example

[453] Road Traffic Act 1988, s143.
[454] Karnow (n 376).
[455] ibid.
[456] Turner (n 205) 115.
[457] Maurice Schellekens, 'No-fault compensation schemes for self-driving vehicles' [2018] 10(2) *Law, Innovation and Technology* 314; David Levy, 'Intelligent no-fault insurance for robots' [2020] 1(1) *Journal of Future Robot Life* 35.

can be seen in New Zealand, in which a no-fault scheme is used for accidents.[458] The classic concern to such a scheme is the lack of a deterrent to both deployers and developers of systems, as there is no direct repercussion to them as an individual,[459] however, there has been no evidence that less caution is taken due to the workings of the New Zealand system.[460] The New Zealand scheme only applies to physical, and some cases of psychological harm, notably not including damage to property and financial loss.[461] In this case, especially when considering AI, several situations would be left without the possibility of compensation, leaving a problematic gap in society. In addition, a no-fault scheme would be significantly costly, raising the question and likelihood of political objections.

In the next part of this review, the literature surrounding data protection are explored, to demonstrate that the current data protection principles when applied to AI are also not sufficient, and to examine the suggestions in relation to this.

2.2 AI and Data Protection

To provide background to RQ2: *To critically examine the impact made by the GDPR on the use and deployment of AI systems, and to what extent would reform of the implemented DPA aid in achieving ethical AI?*

2.2.1 Article 8 ECHR

This section of the literature review explores the relationship between AI and data protection, due to the thesis' scope, and the nature of AI systems having access to data on an unprecedented scale. The lack of human contextualisation available within AI will lead to potential new challenges to the GDPR, and privacy. Article 8 of the ECHR provides the right to respect for private and family life, home and correspondence.[462] Data protection is considered fundamental for safeguarding privacy,[463] and several situations are covered under the ECHR, for example as seen

[458] Accident Compensation Act 2001 (New Zealand).
[459] Laurence Tancredi, 'Designing a no-fault alternative' [1986] 49(2) *Law and Contemporary Problems* 277.
[460] William Gaine, 'No-fault compensation systems' [2003] 326(7397) *BMJ* 997.
[461] Turner (n 205) 105.
[462] ECHR (n 3) Article 8.
[463] European Court of Human Rights, *Guide on Article 8 of the European Convention on Human Rights* (updated 31st August 2020) [184]; *Satakunnan Markkinaporssi Oy v Finland.* App No 931/13 (ECtHR, 27 June 2017) [133].

in the courts, where a public authority stores personal data,[464] or where the use and release of information relating to an individual's private life is stored within a secret register.[465] *S and Marper v the UK* highlighted that in situations where automated processing is used, the need to safeguard individuals is 'all the greater', especially when such data is used for police or state purposes.[466]

Article 8 of the EU Charter provides the right to the protection of personal data,[467] and that such data must be processed fairly for its intended purpose, based on consent of the individual or another legitimate legislative basis. The status of the EU Charter's applicability has been complex post-Brexit; however, it has been argued that the right is enshrined with data protection standards. The GDPR definition of 'personal data' was carried forward from the previous Data Protection Directive (DPD),[468] and is defined as "*any information relating to an identifiable natural person, who can be identified directly or indirectly, in particular reference to an identifier such as a name, identification number, location data, or an online identifier relating to several factors*".[469] The CJEU had also provided clarification on the term in *YS v Minister voor Immigratie,*[470] and rejected that 'legal analysis' which referred to a person fell into the definition.[471] In *Nowak v Data Protection Commissioner,*[472] it was ruled that a handwritten examination script is capable of constituting personal data,[473] which reflects the broad scope of the definition.

In the UK, the ICO is the independent authority that ensures compliance with data protection rules.[474] The GDPR, enforced in 2018 replaced the previous DPD,[475] and directly binds EU MS, but allows for the margin of appreciation. The GDPR applies to

[464] *Rotaru v Romania.* App No 28341/95 (ECtHR, 4 May 2000) [43-44]; *Catt v the UK.* App No 43514/15 (ECtHR, 24 April 2019) [112] and [123].
[465] *Rotaru v Romania* (n 461) [46]; *Leander v Sweden.* App No 9248/81 (ECtHR, 26 March 1987) A/116 [48].
[466] *S. and Marper v the UK.* App No 30562/04 and 30566/04 (ECtHR, 4 December 2008) [2009] 48. E.H.R.R 50 [103].
[467] Charter of Fundamental Rights of the European Union (n 50) Article 8.
[468] Council Directive (EC) 95/46 of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive) [1995] OJ L 281.
[469] GDPR (n 1) Article 4(1).
[470] Joined Cases C-141/12 and C-372/12 *Y.S v Minister voor Immigratie, Integratie en Asiel, and Minister voor Immigratie, Integratie en Asiel v M. and S*. [2014] 62012CJ0141
[471] ibid [1].
[472] Case C-434/16 *Nowak v Data Protection Commissioner* [2017] 62016CJ0434
[473] ibid [51].
[474] Information Commissioners Office, 'Who we are' (ICO Website) <https://ico.org.uk/about-the-ico/who-we-are/> accessed 15th November 2020.
[475] Data Protection Directive (n 465).

‘controllers’; “*any person, public authority, agency or other body that processes personal data of EU citizens*”,[476] and ‘processors’, who process data on behalf of a data controller,[477] and provides protections for all EU ‘data subjects’. The GDPR introduces several safeguards and rights for data subjects, including that information provided by the data controller be given in a concise, transparent, intelligible, and accessible manner,[478] the right of access to personal data, the reasons for it being processed,[479] the right to erasure,[480] and the right to object to their personal data being processed.[481]

The UN comments that new challenges are brought by AI, including to the principles of purpose and use limitation, transparency, and accountability, which are considered the pillars that international data protection is founded upon.[482] AI systems may process personal data through datasets used for training, or through models being applied to personal data, to make inferences relating to individuals.[483] Due to the concept of personal data not being exhaustively defined, Ufert highlights the lack of clarity on whether these inferences of personal data form part of this concept.[484] The EP suggest that the processing of inferences based on personal data is permitted under the GDPR, provided that appropriate safeguards are adopted.[485] There is no further clarification on what is meant by ‘appropriate’, or whether the EP just means compliance to the GDPR mechanisms.

The EDPS comments that due to complex algorithms having the capability to analyse large datasets and make predictions, and the increasing amount of data being collected, monitored and used for ML purposes, challenges to privacy and data protection are presented.[486] The UN recommend that existing data protection

[476] GDPR (n 1) Article 4(7).
[477] ibid Article 4(8).
[478] ibid Article 12.
[479] ibid Article 15.
[480] ibid Article 17.
[481] ibid Article 21.
[482] United Nations (n 82) [35]; United Nations Human Rights Committee, *General Comment No.16: Article 17 (The right to respect of privacy, family, home and correspondence, and protection of honour and reputation)* (Thirty-second Session, 8th April 1988) [10].
[483] European Parliament (n 247) 1.
[484] Ufert (n 4).
[485] European Parliament (n 247) 2.
[486] European Data Protection Supervisor, ‘Artificial Intelligence’ (EDPS website) <https://edps.europa.eu/data-protection/our-work/subjects/artificial-intelligence_en> accessed 15th November 2020.

regulations must be updated to account for these new challenges,[487] which is agreed with by Spyridaki, the Chief Privacy Strategist of SAS Europe, who comments that the GDPR in some cases, complicates the processing of personal data in an AI context.[488] Continuing, this section highlights the challenges brought to data protection by the use and deployment of AI, and identifies the key themes on AI and data protection in the literature.

2.2.2 Article 22 of the GDPR

Article 22 is one of the few provisions in the GDPR that specifically addresses AI technology.[489] The Article provides data subjects the right "*not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects… or similarly significant affects*" the data subject.[490] 'Profiling' is defined under the GDPR as automated processing that uses personal data to analyse or predict aspects relating to "*work performance, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements*".[491]

Article 22 poses a distinct issue for unsupervised AI used for high risk purposes, reflecting the risk-based approach of the GDPR, subjecting those decisions resulting in legal or similarly significant risks to a higher standard of compliance.[492] The Centre for Information Policy Leadership explains that this Article exists to address the concern of handing over complete power to AI, which is likely to produce harmful or dangerous effects.[493] This Article also reflects the inherent concern of AI bias, particularly where there is no human supervision and monitoring, but if AI were to ever reach its true potential, the need for additional rules on the ethical development and application of AI are imperative, to reduce and potentially remove this restriction on high risk systems.

[487] United Nations General Assembly (n 82) [63].
[488] Spyridaki (n 84).
[489] GDPR (n 1) Article 22.
[490] ibid Article 22(1).
[491] ibid Article 4(4).
[492] Centre for Information Policy Leadership, *Artificial intelligence and Data Protection: Delivering Sustainable AI Accountability in Practice, First Report: Artificial Intelligence and Data Protection in Tension* (10th October 2018) 16.
[493] ibid 17.

The question can also be posed to what extent Article 22 applies to decisions that are ‘largely’, rather than ‘solely’ based on automatic processing,[494] potentially reflecting the limited scope of the Article. In addition to this, Privacy International notes that the wording implies that in the absence of decision-making, profiling alone does not give rise to the safeguards established, but does still receive the safeguards contained in Articles 13-15.[495] Articles 13-15 refrain from specifying that decisions need be ‘solely based’ on automatic processing,[496] creating potential inconsistencies and conflicts within the legislation.

The scope of ‘similarly significant’ effects also must be clarified to prevent unnecessary stifling of AI. The Article 29 Working Party comment that the wording is clear, and that only serious impactful effects are included,[497] which fails to provide further clarity. They do, however, comment that the threshold for significance must be based on a similar level to that of legal effects, or that the effect must be sufficiently great or important to be worthy of attention. The term ‘significant’ can be considered as a subjective term, in which effects considered ‘greatly worthy of attention’ may differ between individuals. The Article 29 Working Party do highlight this issue, giving the example of an individual known to be in financial difficulties regularly targeted with adverts for high-interest loans,[498] but they fail to provide further clarification on this, including to what, or who’s standard this Article applies to.

In the same report, the Article 29 Working Party acknowledge that it is difficult to be precise about exactly what would be considered sufficiently significant to meet the threshold for some applications,[499] contradicting their earlier statement. An example is noted on online advertising, which may not always fall under the scope of the Article, but might, depending on the characteristics and extremities of the case,[500] reflecting the difficulties in determining the scope.

[494] Borgesius (n 4).
[495] Privacy International, *Data is Power: Profiling and Automated Decision-Making in* GDPR (April 2018) 10.
[496] ibid 10.
[497] Article 29 Working Party, *WP251 Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679* (February 2018) 21.
[498] ibid 22.
[499] ibid 22.
[500] ibid 22.

Article 22 has a number of exceptions,[501] including where:

- the decision is necessary for entering into a contract,
- where it is authorised by law, to which the data controller is subject to suitable measures to safeguard the data subject's rights,
- or based on the data subject's explicit consent.

The first exception based on entering into a contract is rather broad,[502] and for it to apply, the decisions based solely on automated processing must be 'necessary'. Although the EP give examples of what constitutes necessity,[503] including a high number of cases to be examined, or the capacity for AI to outperform human judgement, the scope may also give rise to abuse.

The second exception includes the requirement of authorisation by law,[504] however, suitable measures must be in place to safeguard data subjects. Recital 71 comments further on these measures,[505] including ensuring that acceptability, accuracy, and reliability are respected.[506] This could also cause tension for the courts and developers, as the term 'suitable' is also subjective. In relation to the first and third exceptions, Article 22(3) implements the same 'suitable measures' standard,[507] interpreted by the ICO as ensuring that data subjects are given information about the processing of their data, that simple ways are introduced for them to request human intervention or challenge a decision, or that regular checks to ensure systems are working as intended are carried out.[508]

The third exception requires the data controller to receive explicit consent from the data subject,[509] which includes the data subject understanding what they are consenting to.[510] Ufert highlights that whilst in theory requiring consent should provide for a sufficient safeguard against fundamental right infringement, it may be more complex to obtain informed consent in situations where AI makes

[501] GDPR (n 1) Article 22(2).
[502] ibid Article 22(2)(a).
[503] European Parliament (n 247) 61.
[504] GDPR (n 1) Article 22(2)(b).
[505] ibid Recital 71.
[506] Article 29 Working Party (n 494) 32.
[507] GDPR (n 1) Article 22(3).
[508] Information Commissioner's Office, *Rights Related to automated decision making including profiling* (July 2020).
[509] GDPR (n 1) Article 22(2)(c).
[510] Article 29 Working Party (n 494) 13.

unpredictable decisions,[511] posing a potential challenge, and restriction, on the use of this exception. Also, the means of obtaining consent has given rise to concern, especially if included in a company's terms and conditions, where it is believed that a low number of citizens, read and understand them in full before accepting. This highlights the need for further safeguards or requirements to ensure consent is gained sufficiently.

In recent years, Article 22 has been subject to the first legal challenge of its kind, in which a trade union on behalf of Uber drivers had filed a complaint in Amsterdam's District Court relating to the use of Uber's alleged 'robo-firing' algorithm.[512] This challenge stemmed from several Uber drivers who were accused of 'fraudulent behaviour' by the app, and automatically fired.[513] It was claimed that Uber had refrained from giving the drivers access to evidence on the matter, and had not allowed them to challenge or appeal their termination.[514] Uber claims that the drivers in this case were only deactivated after manual reviews by a specialist team, and hence the restriction of Article 22 did not apply.[515] This reflects the limited scope of the article and the 'get-out clause' available to businesses or corporations which could be abused.

The court also heard a similar case against Ola (an app comparable to Uber), which is discussed in the section below in the context of the right to an explanation, bringing hope to those who argue the existence of such a right. Continuing, the discussion in academic literature relating to Articles 13-15[516] will be explored, with particular focus and reference to the debated right to an explanation.

2.2.3 Articles 13-15 of the GDPR

Articles 13-15 govern the data subject rights to be informed and right to access data, through ensuring 'fair and transparent processing', by requiring "*meaningful*

[511] Ufert (n 4).
[512] Catherine Wycherley, 'Uber faces landmark GDPR court challenge over alleged firing of drivers by algorithm' (GDPR Report Blog, 29th October 2020) <https://gdpr.report/news/2020/10/29/uber-faces-landmark-gdpr-court-challenge-over-alleged-firing-of-drivers-by-algorithm/> accessed 14th November 2020.
[513] ibid.
[514] ibid.
[515] *Uber v Uber Drivers (Deactivation Case)* [2021] C/13/692003/HA RK 20-302 (Netherlands).
[516] GDPR (n 1) Articles 13-15.

information about the logic involved" be provided for automated decisions,[517] assumed by many academics to be interpreted as a right to explanation.[518] Recital 71, although non-binding, provides further clarification, expressing that data subjects should have "*the right to obtain an explanation of the decision reached after such assessment and to challenge the decision*".[519] The right in the Recital is more clear and direct, in comparison to those included within the Articles, posing the question of why a vaguer option was chosen to be in the binding parts of the regulation, and therefore offering less assurance of such a right.

The EP rightly comment that the use of the word 'logic' included within the Articles is in need of further clarification, and that such 'information' to explain a decision needs to be specified with appropriate examples.[520] One issue surrounding the right to explanation for automated decisions is the consideration whether such an explanation is possible in all situations, as in complex AI systems, it may be impossible to provide an explanation at all. In these cases, it would also be difficult to meaningfully contest, intervene, review, or put an alternative point of view against decisions, giving rise to several issues.[521]

A somewhat simpler solution in this scenario would be prohibiting 'black-box' systems from processing personal data that causes legal or similarly significant effects, however, this would create an avoidable and substantial restriction on the use and potential of AI, resulting in a counter-productive outcome. Further clarification on how this right should be upheld is needed, to ensure the successful progression and future of AI. Uncertainty has also arisen in relation to whether individual explanations should be provided to the data subject. The EP believes data controllers should be put under an obligation to provide this when it is possible and reasonable to do so.[522]

[517] ibid Article 13(2f), Article 14(2g) and Article 15(1h).
[518] Andrew Selbst and Julia Powles, 'Meaningful information and the right to explanation' [2017] 7(4) *International Data Privacy Law* 233; Bryan Casey, Ashkon Farhangi and Roland Vogl, 'Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise' [2019] 34 *Berkeley Technology Law Journal* 143; Edwards and Veale (n 179).
[519] GDPR (n 1) Recital 71(4).
[520] European Parliament (n 247) 4.
[521] Information Commissioner's Office, *How do we ensure individual rights in our AI systems?* (July 2020).
[522] European Parliament (n 247) 63.

The literature debates whether there exists a right to explanation in the GDPR on a substantial level; Goodman and Flaxman argue that the GDPR does create a 'right to explanation', albeit only providing a relatively narrow protection.[523] They also observe that ML systems were "*alone on the spectrum in their lack of interpretability*",[524] acknowledging the issue of the uninterpretable black-box. It also needs to be examined that in some circumstances, an explanation may not be helpful,[525] whether it be that the explanation itself cannot be justified or be understood. Concerning ML systems, the notion of an explanation is referred to by Wachter, Mittelstadt and Russell as needing to be understood by others, and could include "*providing insight into the internal state of the algorithm, or to human-understandable approximations of the algorithm*".[526]

Edwards and Veale doubt the effectiveness of the 'right to an explanation', fearing that it will lead to a transparency fallacy.[527] This is agreed with by Wachter, Mittelstadt and Floridi, who also acknowledge that several algorithmic decisions remain outside of the scope of the GDPR rules,[528] reflecting its limited application in relation to AI. Edwards and Veale further state that the law is "*restrictive, unclear and even paradoxical*" [529] in relation to when any explanation-related right can be engaged,[530] presenting challenges for those who seek to comply. Selbst and Powles somewhat convincingly suggest that the right to explanation should be "*interpreted functionally, flexibly, and should, at a minimum, enable a data subject to exercise their rights under the GDPR and human rights law*",[531] however it may be optimistic for this to be the case in practice.

To address the debate on the right of an explanation within the GDPR, the judgment in Amsterdam against Ola needs to be discussed further. In this case, it was argued that the Article 22 scope did apply to the decisions made, due to decisions being capable of causing legal or similarly significant effects. As drivers received penalties

[523] Goodman and Flaxman (n 195).
[524] ibid.
[525] Edwards and Veale (n 179).
[526] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR [2018] 31(2) *Harvard Journal of Law and Technology* 841.
[527] Edwards and Veale (n 179).
[528] Wachter, Mittelstadt, and Floridi (n 195).
[529] Edwards and Veale (n 179).
[530] ibid.
[531] Selbst and Powles (n 515).

because of the automated decisions, Article 22 was triggered, and Ola was required to explain the logic underpinning the decision.[532] This judgment has been described as surprising, due to the disputed right to an explanation being recognised by the courts, and that the right includes data subjects being able to understand the rationale of the decision itself, rather than just how the system works. This has reportedly been the first instance within Europe to recognise a right to an explanation within the GDPR, and gives a very impactful precedent for the future.

Comparisons and considerations into other 'right to explanations' included outside of the GDPR are also less discussed in the literature, in which an example can be seen in French administrative law, where a right to explanation for algorithmic decisions made about individuals is viewed as broader, and more comprehensive than the GDPR.[533] A similar standard could be considered in a future framework for AI, and could ensure stronger and clearer safeguards against unfair, or unethical decision-making.

2.2.4 Data Protection Principles

Article 5 of the GDPR introduces the data protection principles, including lawfulness, fairness, and transparency,[534] which raise difficult questions in relation to AI and automated decision-making. Article 5 also includes the data minimisation and storage limitation principles, which restrict the amount of data processed for a purpose, and the length it is stored for.[535] The Article 29 Working Party acknowledge the challenge in applying these principles to AI systems.[536] AI benefits from large datasets and retaining data for more efficient and successful training purposes. The scope of the phrase within the data minimisation principle that limits the collection of data to 'what is necessary' needs to be examined in this context, to determine whether it is the most appropriate safeguard for AI. As argued by the Centre for Information Policy Leadership, storing data is inherent to numerous AI systems, especially profiling, and could be considered more advantageous in the sense that

[532] *Ola v Ola Drivers* [2021] C/13/689705/HA RK 20-258 (Netherlands).
[533] Privacy International and Article 19 (n 247) 23; Loi pour une République numérique (Digital Republic Act, 2016-1321).
[534] GDPR (n 1) Article 5(1a).
[535] ibid Article 5(1)(b) and (c).
[536] Article 29 Working Party (n 494) 12.

the more data that is input into AI, the more it can learn from, improving accuracy,[537] and mitigating biased or unfair decisions.

The literature considers how to guarantee algorithms are fair, especially regarding the prevention of biased decisions towards a particular ethnicity, gender or other protected groups.[538] Wachter, Mittelstadt and Russell acknowledge that the best tools for uncovering systematic biases are likely to be based on large-scale statistical analysis, rather than explanations of individual decisions,[539] which provides further justification that the data minimisation principle is inadequate for AI. Counterfactual explanations offer another option; these explanations can provide evidence of whether a decision is affected by a particular protected variable, and therefore where there exists the risk of discrimination.[540] To increase social acceptance and public trust in AI, and to close the current gaps which undermine the trust between data controllers and subjects,[541] counterfactual explanations should be used.[542]

Counterfactual explanations can take the form of a statement of the decision, followed by several statements providing information on what variable(s) would need to change for a more desirable outcome. Counterfactual explanations would therefore allow data subjects to understand why a decision was made, to receive information for use in contesting a decision, and guidance on how their behaviour can adapt for the intended result.[543] Such an idea could reduce the complexity and controversy of the ‘right to explanation’ debate, and if imposed, could provide AI decisions with the capability of being understood simply, as well as clearly identifying situations where protected characteristics are being used unfairly in AI systems.

This concept could also aid in the challenges brought by the notion of transparency, which is lacking in most AI techniques, and therefore making them inherently unexplainable.[544] As explained by Larsson and Heintz, AI transparency needs to be

[537] Centre for Information Policy Leadership (n 489) 17.
[538] Wachter, Mittelstadt and Russell (n 523); Larsson and Heintz (n 182).
[539] Wachter, Mittelstadt and Russell (n 523).
[540] ibid.
[541] ibid; Isak Mendoza and Lee A. Bygrave, ‘The Right Not to Be Subject to Automated Decisions Based on Profiling’ in Synodinou, T., Jougleux, P., Markou, C. and Prastitou, T. (eds) *EU Internet Law: Regulation and Enforcement* (Springer, 2017).
[542] Wachter, Mittelstadt and Russell (n 523).
[543] ibid.
[544] Cambridge Consultants (n 4) 26.

understood in context, and can be seen as both a balancing of interests, and a governance challenge.[545] Transparency is challenged by the 'black-box' nature of AI, whereby technology is simply too difficult or advanced to explain. Especially in cases relating to the use of ANNs, it can be "*practically impossible to explain how information is correlated and weighted in a specific process*".[546] Potential areas that require further legal development in relation to transparency may be on 'algorithmic auditing', or requirements for 'high risk' systems,[547] alike to those introduced by the newly approved AI Act,[548] discussed further in Chapter Five.

2.2.5 Data Protection Impact Assessments

Article 35 of the GDPR imposes Data Protection Impact Assessments (DPIAs) on those systems suggested to be high risk under the regulation, especially new technologies.[549] The inclusion of the phrase 'new technologies' can be assumed to be an inference to AI systems, and therefore a DPIA is needed for companies who use AI to process personal data.[550] The ICO comments that DPIAs should be 'living documents' that are reviewed regularly, and updated when any changes are made.[551] Although DPIAs on the surface seem adequate, Ufert and Wrigley highlight the rising concern of this approach becoming a 'rubber stamping' procedure.[552]

By this, Wrigley comments that the complexity of AI has the potential to be used as another excuse to not fully assess the decisions of systems under the GDPR, and developers could merely let such results be approved by a human, just so that it can be said oversight is included and that the risks have been assessed.[553] It needs to be ensured that this is prevented, which could be achieved within a future framework focused on AI's own particular challenges, to ensure the current and future development of AI can be fully compliant with fundamental rights.[554] Mantelero

[545] Larsson and Heintz (n 182).
[546] The Norwegian Data Protection Authority, *Artificial Intelligence and Privacy* (January 2018) 19.
[547] Casey, Farhangi and Vogl (n 515); European Commission (n 112); Datenethikcommission, *Opinion of the Data Ethics Commission* (Data Ethics Commission, German Federal Ministry of Justice and Consumer Protection, 2019) 144.
[548] Provisional Agreement for the AI Act (n 103).
[549] GDPR (n 1) Article 35.
[550] Information Commissioner's Office (208).
[551] ibid.
[552] Ufert (n 4); Wrigley (n 4).
[553] Wrigley (n 4).
[554] Ufert (n 4).

suggests that in addition to the DPIA, a more detailed human rights, social and ethical impact assessment for such technologies should be mandated.[555] This would not be a technological assessment,[556] but a rights-based and values-oriented model, focused on sectors such as healthcare or crime prevention, rather than the category of technology.[557] An approach like this would encourage the review of societal impact, and would concentrate attention not on the technology, but the context in which the values assume relevance.[558]

2.2.6 Where the GDPR Falls Short

Although the GDPR causes an impact to most AI systems, some systems may be left unaffected. As noted by Privacy International and Article 19, the GDPR provisions on automated decision-making and profiling can be considered crucial, however, data protection frameworks also frequently have exemptions for matters such as national security, allowing the power to limit rights and safeguard in substantial privacy-invasive AI applications.[559] One example of this could be government surveillance, where Cobbe calls for more research and exploration into administrative law, and how it should adapt to more automated forms of decision-making.[560]

In New York, legislation introduced in 2017 establishes a taskforce to examine the city's 'automated decision systems' to make them fairer, and open to more scrutiny.[561] Although the legislation has been described as flawed,[562] the idea for higher standards on administrative use of AI is one to be encouraged. Ofcom note another example of the GDPR providing a barrier or disservice to certain aspects in

[555] Alessandro Mantelero, 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment' [2018] 34(4) *Computer Law and Security Review* 754.
[556] Barbara Skorupinski and Konrad Ott, 'Technology assessment and ethics' [200] 1(2) *Poiesis & Praxis* 95.
[557] Mantelero (n 552).
[558] ibid.
[559] Privacy International and Article 19 (n 247) 23.
[560] Jennifer Cobbe, 'Administrative law and the machines of government: judicial review of automated public-sector decision-making' [2019] 39(4) *Legal Studies* 636; Marion Oswald, 'Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power' [2018] 376 *Philosophical Transactions Royal Society A* 1.
[561] New York Government Website, 'New York City Automated Decisions Systems Task Force' (New York, USA) <https://www1.nyc.gov/site/adstaskforce/index.page> accessed 17th November 2020.
[562] Julia Powles, 'New York City's Bold, Flawed Attempt to Make Algorithms Accountable' (The New Yorker, 21st December 2017) <https://www.newyorker.com/tech/annals-of-technology/new-york-citys-bold-flawed-attempt-to-make-algorithms-accountable> accessed 17th November 2020.

the form of online safety for children.[563] Due to the restriction on the ability to use and collect data sets,[564] it is more difficult to track profiles that share harmful content,[565] providing further reason to explore the potential negative impact of GDPR provisions.

As noted by the EP, the GDPR contains numerous vague clauses, and although the regulation intends to balance data protection and other social and economic interests, there is little guidance provided on how to achieve this goal.[566] The EP propose that the GDPR does not need major changes to address AI, however later acknowledge that several AI-related data-protection issues do not have an explicit answer, which may lead to uncertainties, and costs and unnecessarily restrict the development of AI.[567] These latter arguments reflect the inadequacy of the current data protection framework concerning AI, justifying the need for new regulation, which could be based on the GPDR, to improve on the provisions already laid out.

Also, as noted by Borgesius, data protection regulation only applies to the processing of personal data, and therefore algorithmic decision-making processes can be somewhat outside the scope if the data processed does not relate to or lead to an identifiable person.[568] Therefore, it is argued that there may be a need for a broader scope, to cover all aspects of AI use.[569] Ufert also agrees in this aspect, commenting that specific provisions solely applicable to AI should be adopted to ensure a continuous level of fundamental rights protection.[570] Reflecting on this, decisions that affect individuals must be the priority, but it is important that regulation also considers AI outside of this scope, to promote ethical usage regardless of the impact. The next part of the review focuses on the regulation of AI technology, to highlight the common themes, and areas that need further examination and clarification.

[563] Cambridge Consultants (n 4) 55.
[564] ibid 55.
[565] ibid 55.
[566] European Parliament (n 247) 3 and 7.
[567] ibid 3.
[568] Borgesius (n 4).
[569] ibid.
[570] Ufert (n 4).

2.3 AI and Regulation

To provide background to RQ3: *To critically assess the recent attempts to regulate AI made by the EU and UK, and in reflection, what solutions can be proposed to strengthen safeguards and encourage ethical innovation further?*

2.3.1 The Regulatory Race

Worldwide, countries are racing to be the first to achieve progressive and successful AI regulation, to allow the benefits of AI to be seen in a safe, fair, and positive manner. There has been an emergence of non-binding regulations introduced in the UK to prevent and mitigate existing risks,[571] however it needs to be considered how effective these guidelines are, and if they are making, or able to make sufficient impact. This section continues in reflection of this, taking into consideration the challenges and options of regulation, in addition to the efforts made by the EU,[572] the ICO's guidance[573] and the substantial literature in this area, to justify why further regulation is needed for AI to ensure additional protection to citizen's and their rights.

The weaknesses and gaps that exist within these non-binding and proposed regulations will also be identified, to reflect where this research fits into the academic literature and where it will contribute. Cath and Rodrigues express worry that there has been too much focus on voluntary frameworks for AI, and that the solutions proposed do not go far enough, highlighting the pressing need for tighter technical interpretations of fairness, accountability and transparency.[574] Many companies have adopted self-regulation,[575] however, to ensure the strongest protection, a binding legislative framework is imperative to ensure AI is used and deployed in a fair, consistent, ethical and beneficial manner.

[571] Thilo Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines' [2020] 30 *Minds and Machines* 99.
[572] High-Level Expert Group on Artificial Intelligence (n 100); European Commission (n 112).
[573] Information Commissioner's Office (n 98).
[574] Cath (n 189); Rodrigues (n 242).
[575] Microsoft, 'Principles and Approach (Official website) <https://www.microsoft.com/en-us/ai/principles-and-approach/> accessed 18th November 2020; Google, 'Artificial Intelligence at Google: Our Principles' (official website) <https://ai.google/principles/> accessed 18th November 2020; IBM, 'IBM's Principles for Trust and Transparency' (IBM website) <https://www.ibm.com/policy/wp-content/uploads/2018/06/IBM_Principles_SHORT.V4.3.pdf> accessed 18th November 2020.

2.3.2 Challenges to Regulation

Regulation can be defined as "*the sustained and focused attempt to alter the behaviour of others according to standard or goals, with the intention of producing a broadly identified outcome*".[576] In order for a regulatory regime to work effectively, it must be clear on what it is regulating, and unfortunately, there are numerous definitions of AI amongst experts in the field.[577] This is an issue that needs to be solved if future frameworks are to be introduced. Another issue is the pacing problem, where technological innovation increasingly outpaces legislation. Downes states that "*technology changes exponentially, but social, economic, and legal systems change incrementally*".[578] The new technological capabilities of AI are accelerating the pacing problem, and as a result, will continue to present new challenges to traditional legal rules, requiring the introduction of new governance processes.[579]

The precautionary principle enables decision-makers to adopt precautionary measures when future risks are uncertain and stakes are high.[580] The precautionary principle put simply, is acting with caution. Where emerging technologies carry a risk to society, developers of the technology should bear the burden of proving it will not. If this cannot be proven, the use of such technology should be restricted until it can be shown that it is safe.[581] Those in support of the precautionary principle for AI regulation suggest the best approach is "*better to be safe than sorry*".[582] The EU acknowledges that policymakers deal with complex risks when regulating something

[576] Roger Brownsword and Han Somsen, 'Law, Innovation and Technology: Before We Fast Forward- A Forum for Debate' [2009] 1(1) *Law, Innovation and Technology* 1; Lyria Moses, 'How to Think about Law, Regulation and Technology: Problems with 'Technology' as a Regulatory Target' [2013] 4(1) *Law, Innovation and Technology* 1.
[577] Scherer (n 70).
[578] Larry Downes, *The Laws of Disruption: Harnessing the New Forces that Govern Life and Business in the Digital Age* (Business & Economics, 2009) 2.
[579] Ryan Hagemann, Jennifer Huddleston, Adam Thierer, 'Soft Law for Hard Problems: The Governance of Emerging Technologies in an Uncertain Future' [2018] 17(1) *Colorado Technology Law Journal* 37.
[580] European Parliament Think Tank, 'The Precautionary Principle: Definitions, Applications and Governance' (2015, European Parliament) <https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_IDA(2015)573876> accessed 4th December 2019.
[581] Daniel Castro and Michael McLaughlin, 'Ten Ways the Precautionary Principle Undermines Progress in Artificial Intelligence' (Information Technology and Innovation Foundation, 4th February 2019) <https://itif.org/publications/2019/02/04/ten-ways-precautionary-principle-undermines-progress-artificial-intelligence> accessed 16th November 2020.
[582] Cass Sunstein, 'Beyond the Precautionary Principle' (University of Chicago, 2002) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=307098> accessed 1st July 2022.

that is not precisely calculable in advance.[583] Due to this, the precautionary principle can present opportunities for regulation. As the principle has expanded in scope, it has also grown in profile and authority, notably in Article 174(2) of the EC Treaty,[584] where precaution now acts as a key foundational principle in European Community policymaking.

With the intense race for countries to lead the way to regulate AI, the precautionary approach could hold nations back in ways other countries would not be restricted. Castro and McLaughlin share their concern that focusing on the speculative concerns about AI will limit its development and adoption, and suggest that instead of imposing heavy regulation on AI in anticipation of hypothetical harms, provisions should focus on the current problems, and deal with problems when and if they occur.[585]

Castro and McLaughlin advocate for the 'innovation principle', which argues that the vast majority of emerging technologies pose little risk, and therefore use should be encouraged.[586] The innovation principle advocates a case-by-case approach and stresses the importance of design and structure of regulatory enforcement, to minimise the harm to innovation, whilst still achieving the regulatory goals.[587] The thesis argues for this approach to be combined with the precautionary principle, given the examples that already exist concerning unfair and biased decision-making, which are only likely to increase without effective regulatory intervention, however, with recognition that a case-by-case basis should not be overlooked.

2.3.3 The Options of Regulation

Soft law is a term used to describe non-binding legal instruments, norms or directives explicitly avoiding the imposition of legal obligations on the relevant parties, which have gained traction in international law in the last few decades.[588]

[583] European Commission, *Science for Environment Policy, Future Brief: The Precautionary Principle: Decision-Making under Uncertainty* (Issue 18, 2017) 3.

[584] Consolidated version of the Treaty establishing the European Community [2002] OJ C 325, Article 174(2).

[585] Castro and McLaughlin (n 586).

[586] ibid.

[587] RECIPES Project, *Intra Case Study Analysis* (Funded by EU's Horizon Research, 2020) 72.

[588] Andrei Marmor, 'Soft Law, Authoritative Advice and Non-Binding Agreements' [2019] 39(3) *Oxford Journal of Legal Studies* 507.

The term is argued to be self-contradictory by Senden, who argues that a non-binding law is a contradiction that should not exist.[589] Shelton expresses that soft law often serves as a starting point to hard law or as a supplement to a binding law instrument, and can be used as a 'stepping stone' before a commitment to formal and binding agreements.[590] Abbott and Snidal describe soft law as weak law, and argue that its increasing use has the potential to destabilise and stifle regulatory progress, making soft law no longer serve its purpose.[591]

The Partnership on AI was originally started by several successful businesses that develop and invest in AI, and now also includes professional societies, think tanks, academic AI organisations and charitable groups, such as Amnesty International, UNICEF, and Human Rights Watch.[592] The Partnership aims to develop and share best practices for AI,[593] to address areas including fairness, transparency, privacy, reliability, and the collaboration between AI and society.[594] Marchant expresses hope that the Partnership goes beyond these general principles to produce more robust and specific best practices and guidelines for responsible AI research and applications, to solidify their successful 'soft law' player status.[595]

A form of soft law includes codes of conduct, which are used to uphold principles to establish a universally expected behaviour of those within an organisation, with any breach being investigated as misconduct. Kaptein and Schwartz found that codes of conduct help solidify a good reputation, improve the working climate and level of public trust, as well as prevent unethical and illegal wrongdoing, ensuring behaviour is more predictable.[596] Codes of conduct can be used to build public trust in the roll-out of AI, aiding in correcting public misconceptions. Introducing codes of conduct

[589] Linda Senden, *Soft law in European Community Law* (Hart Publishing, 2004) 109.
[590] Dinah Shelton, D. *Commitment and Compliance: The Role of Non-Binding Norms in the International Legal System* (Oxford University Press, 2000) 10.
[591] Kenneth Abbott and Duncan Snidal, 'Hard and Soft Law in International Governance' [2000] 54 *International Organization* 421.
[592] Partnership on AI, 'Meet the Partners' (official website) <https://www.partnershiponai.org/partners/> accessed 4th February 2020.
[593] ibid.
[594] Partnership on AI, 'About Us (official website) < https://partnershiponai.org/about/> accessed 4th February 2020.
[595] Gary Marchant, '"Soft Law" Governance of Artificial Intelligence' (AI Pulse Blog, 25th January 2019) <https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf?t=po1uh8> accessed 4th February 2020.
[596] Muel Kaptein and Mark Schwartz, 'The Effectiveness of Business Codes: A Critical Examination of Existing Studies and the Development of an Integrated Research Model' [2008] 77 *Journal of Business Ethics* 111.

can be beneficial to respond to media and public pressure and concern, and improve the reputation of the technology.[597]

However, as codes of conduct are a form of soft law, they are not binding, and whether rules are followed remains a matter of internal control and sanctioning.[598] Also, codes of conduct are usually found to contain broad and vague language, and therefore are not specific enough to make the difference needed concerning AI. Drafting codes of conduct is also time-consuming, and conflict could arise if AI is used across sectors which each have their own code.[599] Additionally, codes of conduct need to be developed in light of broader policy and legal changes,[600] so on this basis could complement legislation, but not replace the need for it.

The code of conduct approach has been used for AI by the NHS, who encourage technology companies to meet a 'gold-standard' to protect patient data to the highest benchmarks. [601] The code intends to ease the way suppliers can integrate AI technology into the sector, helping institutions to choose safe and secure technology to improve the services they provide.[602] The code consists of ten principles that outline how the NHS should work with technology companies developing AI systems for use in healthcare.[603] Such an approach can be useful, especially for high risk sectors such as healthcare, however, a binding legislative framework will ensure consistency across sectors, and this thesis favours the argument for enforceable rules.

Another form of soft law governance is in the form of self-regulation, which is used usually in areas where national regulation is lacking.[604] Self-regulation has been used in diverse industries to establish industry standards, and to ensure consumer

[597] Ada-Iuliana Popescu, 'In Brief: Pros and Cons of Corporate Codes of Conduct [2016] 9 *Journal of Public Administration, Finance and Law* 125.
[598] Mark Baker, 'Promises and Platitudes: Towards a New 21st Century Paradigm for Corporate Codes of Conduct' [2007] 23 *Connecticut Journal of International Law* 123.
[599] Popescu (n 607).
[600] Ethics for Artificial Intelligence, 'Problems with Codes of Ethics' (website) <https://www.cs.ox.ac.uk/efai/developing-codes-of-ethics-for-ai/downsides-of-codes-of-ethics/> accessed 5th February 2020.
[601] Department of Health and Social Care (n 51).
[602] ibid.
[603] ibid.
[604] Virginia Haufler, *A Public Role for the Private Sector: Industry Self-Regulation in a Global Economy* (Carnegie Endowment for International Peace, 2001) 8.

confidence.[605] Businesses commonly use self-regulation to increase public trust and reputation, particularly where there is the absence of government intervention or the threat of excessive regulation.[606]

The EP notes that although self-regulation can offer a starting point for benchmarks, regulatory frameworks are needed to ensure compliance with fundamental rights.[607] However, the monitoring and remediation processes can be more effective under self-regulation, meaning consumers are protected sooner, which could boost public confidence in AI at a quicker pace than general legislation.[608] Self-regulation allows for incremental and radical innovation, and flexibility allows for more experimental testing rules, knowing that they can easily and quickly be removed.[609]

In this respect, soft law could offer advantages due to its flexibility.[610] However, soft law would not be enough, and a binding regulatory framework is a necessity to set a foundation, upon which soft-law approaches can be established to build on the foundation. It is argued that open and structured discussion, followed by strong and clear regulation is needed for AI.[611] Castelvecci believes that self-regulation would not work, due to the lack of binding status.[612] There is also a common perception that self-regulation allows rules to be made less in the public interest and more to protect other interests, which would be ineffective in promoting ethical AI that upholds human right standards.[613] Also, self-regulation may have a negative impact on regulatory uncertainty, which may cause businesses to delay investment decisions, leading to a stifling of innovation.[614] Hadfield also believes that regulatory oversight is necessary, and that discussions surrounding soft law approaches diminish the complexity of the issue.[615]

[605] Ethics for Artificial Intelligence (n 610).

[606] J. Boddewyn, 'Advertising Self-Regulation: Private Government and Agent of Public Policy' [1985] 4(1) *Journal of Public Policy & Marketing* 129.

[607] European Parliament (n 375); European Economic and Social Committee opinion on AI, 'The Consequence of AI on the digital single market, production, consumption, employment and society' [2017] 31(8) *Official Journal of the European Union* 1.

[608] Castro (n 57) 5.

[609] Boddewyn (n 616).

[610] Tarelli (n 53).

[611] Castelvecci (n 55).

[612] ibid.

[613] Castro (n 57) 9.

[614] ibid 9.

[615] Hadfield (n 54).

2.3.4 Ethical Regulation

Adopted in 2019, the OECD introduced five principles for AI, to promote systems that are innovative, trustworthy and respect human rights.[616] These principles include common themes already identified, such as transparency and responsibility, and the need for accountability mechanisms.[617] Although not legally-binding, these principles were highly influential, and were consequently backed by the EC, who has since developed further guidance based on these principles. One of the most prominent reports released in respect of this was the Guidelines for Trustworthy AI, which clearly stated that trustworthy AI must be lawful, ethical, and robust, and provided seven key requirements that AI must meet to be deemed trustworthy.[618]

The guidelines define AI as "*systems that display intelligent behaviour by analysing their environment and taking actions- with some degree of autonomy- to achieve specific goals*".[619] The guidelines covers the key themes consistent in the literature, including privacy,[620] fairness,[621] accountability,[622] safety,[623] human control,[624] and transparency, explainability and interpretability.[625] The guidelines however, lack in several areas that could have also been considered, including sustainability, social cohesion, job displacement, responsible research funding, public awareness and education, and certification procedures for AI. These guidelines were originally welcomed by the literature due to the intention of creating a consistent approach across the EU,[626] but soon generated debate. The final section looks to 'potential longer-term concerns' of AGI,[627] which created debate of whether AGI should even be considered when it is not clear whether it will be achieved. In the Draft Guidelines,

[616] OECD, *Recommendation of the Council on Artificial Intelligence* (Legal Instrument 0049, adopted May 2019, Amended November 2023) 7-8.
[617] ibid 8.
[618] High-Level Expert Group on Artificial Intelligence (n 100) 8.
[619] ibid 3.
[620] ibid 17.
[621] ibid 18.
[622] ibid 19.
[623] ibid 16.
[624] ibid 15.
[625] ibid 18.
[626] Nathalie Smuha, 'The EU Approach to Ethics Guidelines For Trustworthy Artificial Intelligence' [2019] 20(4) *Computer Law Review International* 97; Insurance Europe, *Response to the consultation on draft guidelines for trustworthy artificial intelligence (AI)* (Position Paper, February 2019) 1.
[627] High-Level Expert Group on Artificial Intelligence (n 100) 35.

this is noted as a highly controversial topic with the authors themselves, describing their concerns as 'speculative'.[628]

Another key debate on the Guidelines surrounds the 'tone' the report is written in, viewed by some as too negative,[629] and others as too optimistic.[630] In relation to respect for democracy, justice, and the rule of law, the guidelines suggest that AI interferes with democratic processes and "*undermines the plurality of values and life choices*".[631] This is one of many interpretations which has been argued to suggest AI as inherently untrustworthy, those arguing that this is not supported by evidence, and can diminish public acceptance of AI.[632] Identified in the consultation, many claimed that in some instances, there was a conflation between ethics and law, creating the possibility of confusion, and potentially suggesting an 'ethics-washing' approach.[633]

This is emphasised in Insurance Europe's position paper, where they express that the guidance lacks an adequate level of connection between the different layers of ethics and law, and calls for further clarification.[634] The guidelines are also said to fall short in addressing global competition,[635] which should be considered given the current regulatory race.[636]

The non-binding nature of these guidelines also give rise to the concern regarding the lack of compliance.[637] Access Now acknowledge that although the recommendations take a step forward, they fall short in enforcing the highest standard of human rights compliance, and somewhat bravely call for "*concrete, actionable policies instead of the promotion of so-called trustworthy AI*".[638] This has

[628] European Commission, *Draft Ethics Guidelines for Trustworthy AI* (High-Level Expert Group on Artificial Intelligence, 18th December 2018) 12.
[629] Smuha (n 636); Center for Data Innovation, *Recommendations to the EU High Level Expert Group on Artificial Guidelines for Trustworthy AI* (2019) 1.
[630] Smuha (n 636).
[631] High-Level Expert Group on Artificial Intelligence (n 100) 13; Center for Data Innovation (n 639) 1.
[632] Center for Data Innovation (n 639) 2.
[633] Ben Wagner, '*Ethics as an escape from regulation'* in Bayamlioğlu,E. Baraliuc, I. Janssens, L. (eds) *Being Profiled* (Amsterdam University Press, 2018); Michael Veale, 'A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence [2020] 11(1) *European Journal of Risk Regulation* 1.
[634] Insurance Europe (n 636) 2.
[635] Center for Data Innovation (n 639) 2.
[636] ibid 2.
[637] Smuha (n 636); Insurance Europe (n 636); Veale (n 643).
[638] Fanny Hidvegi and Daniel Leufer, 'European Union: more big words on AI, but where are the actions?' (Press Release, 26th June 2019) < https://www.accessnow.org/european-union-more-big-words-on-ai-but-where-are-the-actions/> accessed 19th November 2020.

been a reoccurring argument from Access Now, who stress that that voluntary ethical guidelines must be only the first step in addressing the concerns and issues posed by AI technology.[639]

Toh agrees, particularly in light of technologies such as FRT, and proposes the model of Oakland's surveillance oversight law.[640] Oakland's law requires government agencies to provide public documentations on the details and objectives of the use of AI, in addition to putting in place safeguards relating to the collection of data, regular audits, with approval needed before technology is adopted.[641] This model reflects a promising approach, in particular due to the additional focus on data protection, and although specified only for government use in this case, could be developed on a broader and wider basis. Nemitz pushes the argument against the non-binding guidelines further, suggesting that failing to regulate AI by hard law would effectively amount to the end of democracy.[642]

2.3.5 White Paper on AI

The White Paper on AI, introduced in February 2020, was one of the first significant documents of its kind proposed by a global power.[643] Narayanan acknowledges that the GDPR began itself as a White Paper, reflecting the potential significance of the document.[644] Somewhat worryingly, the White Paper included several definitions of AI, and the majority response was negative, especially in relation to its arguably limited scope.[645] The White Paper proposed mandatory requirements only on systems considered to be 'high risk', to ensure that regulatory intervention was focused and proportionate.[646]

[639] ibid.
[640] Toh (n 287).
[641] Oakland Municipal Code, Ordinance adding Chapter 9.64 Establishing Rules for the City's Acquisition and Use of Surveillance Equipment 2018 (Oakland, US)
[642] Paul Nemitz, 'Constitutional democracy and technology in the age of artificial intelligence' [2018] 376 *Philosophical Transactions Royal Society A* 1.
[643] European Commission (n 112).
[644] Narayanan (n 271).
[645] European Data Protection Supervisor (n 114) 7 and 22; Leufer and Jakubowska (n 242); Hidvegi and Leufer (n 242); Daniel Castro and Eline Chivot, 'How the EU should revised its white paper before its published' (Center for Data Innovation, 1st February 2020) <https://datainnovation.org/2020/02/how-the-eu-should-revise-its-ai-white-paper-before-it-is-published/> accessed 4th December 2020.
[646] European Commission (n 112) 18.

The EC acknowledged that the differentiation between 'high risk' and 'low risk' systems needed to be clear for all parties concerned.[647] This differentiation began by categorising high risk as being subject to two cumulative criteria; firstly, that AI is deployed in a sector where significant risks can be expected, and secondly, that it be used in a manner that significant risks are likely to arise.[648] The EC argued that the application of this cumulative criteria ensures that the scope of future regulatory framework would be targeted, and provide legal certainty,[649] however, this is largely disagreed with in the literature,[650] and was viewed as too optimistic. This two-fold criteria has been adapted with the progression of the regulation,[651] which is analysed in further detail within Chapter Five.

The criticism of the criteria focused on the vague term of 'significant risk'. An example already discussed in this Chapter reveals the issues in applying this term, whether this would cover an individual known to be in financial difficulties regularly targeted with adverts for high-interest loans,[652] which arguably would amount to a significant risk. In a situation where the individual is not in financial difficulties, the level of risk would not be significant. To address this, Access Now advocate for a burden of proof approach on developers to show that systems do not meet the threshold of risk, but instead of referring to risk, suggests that human rights violations should be the focus.[653] This could be achieved through a mandatory Fundamental Rights Impact Assessment (FRIA) for AI in every domain, which is made publicly accessible and open to challenge.[654] In addition to mandatory assessments, Access Now call for a legal framework to be developed that prohibits systems which violate human right standards,[655] which coincides with the thesis' argument, that the priority of regulation should be safeguarding human rights.

The White Paper's lack of rules concerning public sector use of AI technology was also subject to criticism. It is disappointing to note the several 'sensible' proposals

[647] ibid 17.
[648] ibid 17.
[649] ibid 17.
[650] European Data Protection Supervisor (n 114) 11; Leufer and Jakubowska (n 242); Hidvegi and Leufer (n 242).
[651] Provisional Agreement for the AI Act (n 103), Article 6.
[652] Article 29 Working Party (n 494) 22.
[653] Leufer and Jakubowska (n 242).
[654] ibid.
[655] ibid.

that were in the original draft but not included in the final draft, including a moratorium of FRT, and special rules for the public sector.[656] The White Paper calls for a 'broad European debate' on FRT,[657] a somewhat lighter approach than originally planned. This emphasises a concern highlighted in the UN's report, which argues that new digital technologies are deteriorating the interaction between the state and the most vulnerable in society, and that additional and stronger safeguards are necessary.[658]

This is further emphasised by the recent case of *R(Bridges)*[659] where police use of FRT was found to be unlawful, and a case in the Netherlands where a surveillance system for detecting welfare fraud was found to violate basic human rights.[660] MacCarthy and Propp acknowledge that although biometric information for use in FRT is already subject to existing legislation, including the GDPR, the technology can vary considerably depending on the context of its use.[661] Therefore, future legislative frameworks need to consider the context of use, to ensure the fullest protection of individual rights and freedoms. Consistency in safeguards is needed both in the public and private sectors, and Digital Europe advocate for the introduction of such safeguards to be matched in equivalence with boosting research funding and education in relation to AI.[662] They also highlight the importance of future legislative frameworks to consider alignment with other existing regulations which already apply.[663]

2.3.6 ICO's Guidance on AI Auditing Framework

The ICO released their guidance on the AI auditing framework in 2020.[664] In the guidance, AI is referred to as an umbrella term used for a range of technologies that mimic human thought.[665] The guidance provides a clear methodology to audit AI

[656] European Commission (n 111) 14-15.
[657] European Commission (n 112) 22.
[658] United Nations, *Extreme Poverty and Human Rights* (Seventy-fourth session, 11th October 2019) 16.
[659] *R(Bridges) v CC of South Wales Police* (n 49).
[660] *Dutch Legal Committee for Human Rights v State of the Netherlands* (n 319).
[661] Mark MacCarthy and Kenneth Propp, 'The EU's White Paper on AI: A Thoughtful and Balanced Way Forward' (LawFare Blog, 5th March 2020) <https://www.lawfareblog.com/eus-white-paper-ai-thoughtful-and-balanced-way-forward> accessed 3rd December 2020.
[662] Digital Europe, *Response to the European Commission's AI White Paper Consultation* (17th June 2020).
[663] ibid.
[664] Information Commissioner's Office (n 98).
[665] ibid 7.

applications to ensure that personal data is processed fairly.[666] The guidance focuses on data protection, and highlights the importance of accountability and governance of AI, including DPIAs,[667] fair, lawful and transparent processing,[668] data minimisation and security,[669] as well as rights relating to automated decision-making, and compliance with such rights.[670]

Several academics welcomed the guidelines, although expressed confusion on whether the guidance was on best practice, the interpretation of data protection legislation, or both.[671] The ICO responded, clarifying that the guidance amounts to best practice for data-protection compliant machines, and acknowledged the need for this to be made clearer in future publications.[672] On consultation, several respondents questioned the lack of guidance on specific individual rights,[673] and how they should be protected. This highlights the pressing need for clarity on the alignment between AI guidance and legislative frameworks, such as the GDPR.

Kazim and Koshiyama call for more explicit discussion between data protection and how it translates into auditing AI, to make clear whether data protection regulation is adequate, or whether it needs to be adjusted.[674] The guidance provided by the ICO fails to add clarity to this, and it is not clear whether data protection related rights and AI impact related rights are parallel to each other, or if there is need for a more detailed assessment.[675] McAuley, Koene and Chen agree with this, expressing that there is a 'general disconnection' between the title of the ICO's guidance, and the content showing compliance with the GDPR.[676] In the ICO guidance, it is suggested that using AI systems may not be the most appropriate option, or in some cases should not be considered an option for use by organisations at all.[677]

[666] ibid 1.
[667] ibid 12.
[668] ibid 36.
[669] ibid 64.
[670] ibid 86.
[671] Information Commissioner's Office, *Summary of response to the consultation on ICO guidance on the AI auditing framework, with comments* (2020) 5.
[672] ibid 5.
[673] ibid 8.
[674] Kazim and Koshiyama (n 123).
[675] ibid.
[676] Derek McAuley, Ansgar Joene and Jiahong Chen, *Response to ICO consultation on the draft AI auditing framework guidance for organisations* (Horizon, 1st May 2020) 3.
[677] Information Commissioner's Office (n 98) 39-40.

McAuley, Koene and Chen call for further guidance in relation to this aspect, suggesting that common technical and commercial factors should be taken into account before an organisation decides to adopt AI technology.[678] In reference to the concern above, as expressed by the Centre for Information Policy Leadership, that although reassessing the scope of the data protection principles in the context of AI is not a simple task, it is essential to avoid unnecessary restrictive regulatory requirements.[679] As highlighted by Rodrigues, privacy and data protection measures are only effective when properly applied, and when sufficiently monitored and enforced,[680] reflecting how mandatory AI legislation could be key in achieving this.

2.4 Conclusion

This chapter has demonstrated the breadth of academic discourse surrounding the interaction of AI with human rights, liability frameworks, data protection laws and regulatory challenges, but significant gaps remain. The first section of the literature review (2.1.1) shows that whilst AI has the potential to benefit society, it poses substantial ethical and human right challenges, including issues such as bias, fairness and transparency. The literature calls for more robust protection of human rights, highlighting that current approaches lack the adequate mechanisms to address these risks. This section of the discussion informs the thesis' focus on RQ1, which seeks to examine how AI's impact on human rights can be mitigated through a new regulatory framework. The major themes which emerged in the literature of bias, fairness and transparency are discussed further in Chapter Three, which also includes recommendations to address these matters.

After reflection on the second part of the literature review (2.1.2), it is evident that several options exist to address AI and liability, each with their own advantages and drawbacks. Although various liability options are proposed in the literature, no consensus exists on the best approach. This section of the review reveals a clear gap in the law being able to handle AI's unpredictable nature, signalling the need for new rulings and guidance. If AI continues to develop and progress at its current rate, the concern towards the foreseeability of machines will only increase and become

[678] McAuley, Joene and Chen (n 686) 4.
[679] Centre for Information Policy Leadership (n 489) 15.
[680] Rodrigues (n 242).

more complex, making it imperative that the scope of future liability rules can deal with this.[681] This supports the thesis' argument that current liability regimes are insufficient, which is discussed further in Chapter Three, where a more comprehensive liability option is proposed.

It is evident in the third section of the literature review (2.2) that the GDPR has a substantial impact on AI applications and automated decision-making. Whilst the GDPR represents a significant regulatory effort, the literature suggests that the provisions are inadequate to address the challenges brought by AI. This leads to a key observation that current data protection laws are inadequate for AI's unique risks, particularly regarding automated decision-making under Article 22. The findings from this section form the foundation of analysis within Chapter Four, which addressed RQ2 and evaluates whether reform to the DPA is necessary to promote a human rights centred approach to AI regulation.

The final section of the literature review (2.3) highlights the challenges and options of producing a framework for AI, with hard legislation being the most effective choice to ensure compliance and enforcement of safeguards. The absence of a standardised definition of AI, and a lack of focus on human rights in the forefront of current proposals are recurring concerns in the literature. These gaps are addressed in Chapter Five to address RQ3, where the thesis argues for a more cohesive and human-centric regulatory framework. The major themes which emerged in the literature are assessed further in Chapters Three and Four. In addition, Chapter Five uses this groundwork discussion of regulatory attempts to examine the most recent regulatory developments in the area, highlighting where clarity is still lacking.

The literature reveals that whilst steps have been taken to regulate AI, gaps remain in areas related to the prioritisation of human rights, challenges to current liability frameworks and when applying data protection provisions. The thesis intends to fill these gaps by proposing a comprehensive regulatory framework that address these points, with recommendations to ensure AI systems are not only legally compliant, but also ethically developed and deployed. The following chapters will build on the

[681] European Commission (n 112) 10.

themes identified in this Chapter, offering proposals for regulatory advancements that align with the evolving complexities of AI systems.

Chapter 3: AI Ethics and Accountability

RQ1: How do AI systems challenge ethics, human rights, and current liability rules, and how would a new framework address these issues?

3.1 Introduction

The purpose of this chapter is to explore the challenges identified in the literature relating to the human rights implications and liability issues concerning AI technology. This discussion builds upon two key themes introduced in the prior chapter by directly addressing the challenges to human rights and liability regimes. These areas are the foundation for the overarching argument that underpins the thesis: the need for comprehensive, human rights-centred regulation. This chapter follows on from the exploration on these areas in Chapter Two to make proposals in order to address the challenges identified. This chapter uses a doctrinal approach to highlight how a proposed new framework for AI could be beneficial in addressing the impact on human rights and the challenge to liability principles. This chapter contributes to the literature by providing recommendations to not only strengthen protections for citizens' fundamental rights and freedoms, but to clarify rules for those who are developing, deploying, and using AI technology.

Through the discussion of human rights and the challenges to liability posed by AI, this chapter contributes to the thesis by establishing a clear foundation for the regulatory proposals introduced in later sections. The chapter also reinforces the broader argument that AI regulation must be forward looking, and be human-centric. As noted in the literature review, AI technology evidently brings substantial challenges to human rights, and major concern exists in relation to the potential of bias in systems. It is inevitable that given the current rate of innovation and progression of AI, eventually, focused regulation will need to be put in place, and such regulation must have a human rights centred approach to promote and encourage ethical AI.

To begin, this chapter explores the challenges relating to the **accuracy**, **fairness,** and **bias** in machines, and provides suggestions on how these challenges can be addressed within a new proposed framework for AI. Within the GDPR,[682] specified data protection principles are able to emphasise the most important aspects that must be considered, and make clear the ethical basis of the regulation. Under Article 5 of the GDPR, the data protection principles that data controllers must adhere to include lawfulness, fairness and transparency, purpose and storage limitation, data minimisation, accuracy, and integrity and confidentiality.[683] A similar approach is advocated for within the thesis' proposed recommendations for a new framework for AI, whereby such principles are adapted to focus on the specific challenges posed by AI machines, with the aim of providing further clarity and security. In addition, the proposed recommendation includes an obligation for those, including data controllers, to comply with monitoring requirements, which stems from the need to be able to identify patterns in systems before having the ability to address any anomalies. It is essential that such monitoring requirements are binding, to ensure full adherence and the ability to enforce requirements, and apply to all AI systems that process personal data, to ensure sufficient protection to citizens, and to encourage the progression of ethical AI.

In addition, the concepts of **transparency**, **explainability** and **interpretability** in consideration of AI are assessed, and proposed as a potential requirement for AI systems. For AI to be regulated sufficiently, the workings of how a decision is made need to be understood to a certain degree, to provide the opportunity to ensure such processes are working to the appropriate standard. The proposed binding principles would ensure the future of AI systems are developed with ethical practices at the forefront, and would aid in matters where decisions are challenged. The second section of this chapter further examines the liability issues that arise when using AI systems. In reflection of the literature review, the current liability frameworks are insufficient in addressing the complexities of AI, highlighting the need for further guidance, or additional rules to address these matters. The recommendations suggested combine those within the literature, and aim to provide a new perspective on how liability for AI could be addressed.

[682] GDPR (n 1).
[683] ibid, Article 5.

3.2 Human Rights Implications

3.2.1 Accuracy, Fairness and Bias in AI Systems

The progression of AI technology has the capability of causing potentially devastating and long-lasting impacts on human rights, if not developed and used with ethical practices in mind, and more importantly, if not legally required to. AI systems are now widely used by the state and major corporations, affecting the public on an everyday basis, essentially making it imperative for binding rules to ensure any interference with human rights is minimised. In respect of this, a proposed new framework for AI could aid in ensuring the ethical development and use of AI systems by requiring compliance with rules relating to **accuracy**, **fairness,** and **bias**.

It is essential that AI systems, especially if in public use, are **accurate**, **fair,** and **non-discriminatory**, to sufficiently safeguard citizens and their rights, and to promote the ethical progression of AI. Given the nature of AI technology, it is inevitable that some cases will arise where AI acts in an unpredictable way, which consequently may cause a negative impact to an individual's rights. This arguably would predominantly include infringements to the right to privacy,[684] the freedom of expression,[685] and discrimination of protected characteristics under the Equality Act,[686] but could also affect many other fundamental rights.[687]

To identify any unpredictable decisions made, and to prevent future occurrences of unpredictability, AI systems need to be developed with a sufficient level of transparency, so the workings of AI can be reviewed and assessed where necessary. In addition to the proposed standard for **transparency** included within the thesis' recommendations for a new framework for AI, it is also imperative that consistent and efficient options of redress are available and accessible, to uphold fundamental right protections. As reflected in the GDPR,[688] for redress to be available, the concepts of **explainability** and **interpretability** are essential, to allow

[684] ECHR (n 3) Article 8.
[685] ibid Article 10.
[686] Equality Act 2010, s4.
[687] ECHR (n 3); Charter of Fundamental Rights of the European Union (n 50).
[688] GDPR (n 1) Articles 13-15 and Article 22.

AI decisions and decision-making processes be understood and challenged by laymen.

Ensuring accuracy, fairness, non-discrimination, transparency and explainability in AI systems aligns with the safeguarding of fundamental human rights, particularly in regard to access to justice and the rule of law. Through requirements to ensure AI systems be accurate, transparent, explainable and fair, a proposed framework would not only satisfy regulatory and technical needs, but would also uphold core fundamental human rights through empowering individuals to understand, challenge, and seek redress for AI-driven decisions that may affect them.

AI Accuracy

For AI systems to be effective and achieve their intended aim, it is essential that the decisions made by the related technology are accurate. Accuracy in AI can be defined as the total number of correct predictions made within the decision-making process, in reference to the number of false positives and false negatives.[689] False positives refer to instances where a positive outcome is produced, but predicted incorrectly by the ML algorithm, whereas false negatives are negative outcomes that are predicted incorrectly by an algorithm. For example, if an image recognition system that alerts the user to identifiable humans, identifies an image as a person incorrectly, and sends an alert, this would be considered a false positive. Whereas the machine failing to recognise a human, and hence, failing to send an alert, would be considered a false negative.

These concepts have been discussed in the *Republic of Poland* case,[690] whereby the Attorney General (AG) highlighted the importance for the error rate of false positives in systems to be kept as minimal as possible.[691] In circumstances where this was not possible, such systems should be prohibited to ensure those affected are safeguarded effectively. This has also been reflected in the *Ligue des droits*

[689] Ashish Mehta, 'What is Accuracy, Precision, Recall and F1 score? What is its significance in Machine Learning?' (Medium Blog, 17th September 2020) <https://medium.com/ai-in-plain-english/what-is-accuracy-precision-recall-and-f1-score-what-is-its-significance-in-machine-learning-77d262952287> accessed 22nd January 2021.

[690] Case C-401/19 *Poland v Parliament and Council* [2022] OJ C 270.

[691] ibid, opinion of AG, [214].

humains case, where remarks were made on the "*fairly large number of false positive results"* ,[692] being a factor behind the decision that the system should not be used, to ensure the proper safeguards to individual rights.

In agreement with the AG in *Republic of Poland[693]* case, it is essential that AI machines have a high standard of accuracy, particularly when in public use, to safeguard those who are subject to the decision-making. 'Accuracy' is also listed as one of the key requirements in the EC's Guidelines for Trustworthy AI,[694] and is accepted as a necessity to build public trust. Requirements for **accuracy** should be set out in legislative provisions, and may have to include parameters or thresholds based on what can be deemed reasonably conceivable.[695] The main question and difficulty that arises in this context is how to define a 'high standard' of accuracy in the context of AI. Defining a 'high standard' of accuracy is subjective, and could depend on several factors, including the significance of decisions being made, and the uses of the system.

This thesis rejects the idea that accuracy in AI is good enough if it is better than a human counterpart. Systems have evidenced this in several different fields, having both higher levels of accuracy and taking less time to make decisions in comparison to humans.[696] In this context, the advantages of AI reflect these values of being able to make more accurate and quicker decisions in comparison to humans, and hence, this should be a pre-requisite requirement for deploying machines rather than something to 'aspire' for.

Several academics in the literature refer to numbered percentages in defining 'good accuracy', with reports of 80% becoming an accepted norm,[697] yet others stating that

[692] Case C-817/19 *Ligue des droits humains v Conseil des ministers* [2020] OJ C 36/16.
[693] *Poland v Parliament and Council* (n 700).
[694] High-Level Expert Group on Artificial Intelligence (n 100) 14.
[695] *Poland v Parliament and Council* (n 700), opinion of AG [211].
[696] Marc Lanovaz and Kieve Hranchuk, 'Machine Learning to Analyze Single-Case Graphs: A Comparison to Visual Inspection' [2021] 54(4) *Journal of Applied Behavior Analysis* 1541; Stephen Weng, Jenna Reps, Joe Kai, Jonathan Garibaldi and Nadeem Qureshi, 'Can Machine-Learning Improve Cardiovascular Risk Prediction using Routine Clinical Data? [2017] 12(4) *PLoS One* 1.
[697] Charles E. Olson, Jr. 'Is 80% accuracy good enough?' (Senior Image Analyst and Michigan Tech Research Institute, 2008) <https://www.asprs.org/a/publications/proceedings/pecora17/0026.pdf> accessed 13th February 2019.

anything greater than 70% is 'great model performance'.[698] The challenge of setting standards according to a numbered calculation takes away the significance of the discussion in the context of the inaccuracies.[699] For example, a system which has proven 95% accuracy may seem 'impressive' in the context of the above discussion, however, if this system was used in the medical field to identify cancerous cells, the 5% inaccuracy could lead to significant and severe consequences.

In response to this discussion, it may seem obvious that the only reasonable standard is to expect 100% accuracy, to ensure the strongest safeguards to individuals. However, it is highly questionable whether 100% accuracy can actually be achieved, and deemed by many to be impossible.[700] Also, due to those systems which use ML to increase their datasets based on the previous learning experiences, the percentage of accuracy would be forever changing, not giving a clear indication of how accurate the machine really is. For this reason, it must be emphasised that whilst accuracy in machines is imperative, it must be required in line with several other safeguards to ensure the upmost protection to individuals.

Regarding RQ1, it is evident that possible inaccuracies in AI systems, particularly if left unaddressed, could cause a potentially devastating impact on human rights. This could cause significant reputational damage to governmental and private sector corporations if a heavy reliance was placed on AI systems which lack a high level of accuracy. If these systems were to go wrong, with a lower level of accuracy, there is a higher risk of mistakes being made, and corporations could suffer not only damage to their reputation, but also revenue losses, and diminished public trust in their services.[701] Depending on the machine in question and the context of its use, these risks may lead to more severe consequences. For example, if an algorithm which

[698] Kirsten Barkved, 'How to know if your machine learning model has good performance' (Obviously AI, 9th March 2022) <https://www.obviously.ai/post/machine-learning-model-performance> accessed 20th May 2022.

[699] Kemal Tugrul, 'When Even a Human is Not Good Enough as Artificial Intelligence' (Towards Data Science, 6th May 2018) <https://towardsdatascience.com/when-even-a-human-is-not-good-enough-as-artificial-intelligence-c39c9fda4644> accessed 20th May 2022.

[700] Alex Castrounis, *AI for People and Business: A Framework for Better Human Experiences and Business Success* (O'Reilly Media, 2019) 174; Fabian Sterken, 'AI isn't 100%' (Indica, 28th June 2022) <https://indica.nl/blog/2022/28/6/ai-artificial-intelligence> accessed 29th June 2022.

[701] Charlie Pownall, 'Understanding the reputational risks of AI' (CPC & Associates, AI Trust & Transparency Project, 2019) <https://www.researchgate.net/publication/340088726_Understanding_the_Reputational_Risks_of_AI> accessed 20th March 2021.

lacks a high level of accuracy is used within the medical field, this could lead to human injury or loss of life, posing a clear issue to be addressed.

AI accuracy is critical for systems used throughout society, easily seen in the extreme example of healthcare, in which a robotic aid used to assist with surgeries could cause destructive consequences. In addition to this, several AI systems exist and are in widespread use that produce decisions that substantially affect individuals, and consequentially their rights. For this reason, to promote and encourage ethical AI, future regulation must enforce a high standard of accuracy in available AI systems, with a baseline given in terms of percentage, but with further guidance provided in terms of what is needed for sector-specific and particular uses of systems before their deployment. This is supported within the White Paper for AI, which also acknowledges that future regulation should include accuracy requirements, and that a specified benchmark should be provided.[702]

Setting a benchmark for accuracy is not an easy task, and needs to be considered in the context and consequences of the machine's decision-making. For example, it can be argued that a healthcare diagnosis algorithm has a higher need for accuracy, in comparison to online targeted advertising, due to how the output of the algorithm may be used. From this assumption, it is easy to see why the risk-based approach to legislation has been considered and arguably favoured by Parliamentarians when setting standards. It also needs to be considered whether the accuracy rate of algorithms sets a higher standard than its human counterpart, to ensure that systems are being used in areas to aid and exceed human abilities, rather than being used for the sake of trialling and expediting technology.

This element is drawn upon by Aizenberg and Van Den Hoven, who state that systems should be categorised into goal-directed or practice-directed, and that human benefits, such as building relationships and sharing emotions, should not be ignored and pushed to one side for the sake of implementing technology.[703] Focusing on the accuracy element, Enlitic, who are known as a pioneer in medical

[702] European Commission (n 112) 18.
[703] Evgeni Aizenberg and Jeroen Van Den Hoven, 'Designing for Human Rights in AI' [2020] 7(2) *Big Data & Society* 1.

ML,[704] have developed a tool to aid radiologists in identifying diseases and medical conditions, and through rounds of testing, has revealed a greater accuracy rate than three combined radiologists by 50%, and a 0% false-negative rate overall.[705] This system has proven a clear advantage to its human counterparts, and particularly due to its minimal false-negative rate, reflects the ability for both radiologists and patients to be able to trust the machine and its decision-making abilities.

The importance of accuracy in machines can be assessed in two stages, in a pre-emptive nature before such machines are deployed and used, and over time once such technology is made available and is in use. The European Union Agency for Fundamental Rights acknowledge that although accuracy within AI is generally improving, the risk of errors remains real, in particularly for minority groups.[706] Accuracy is included as one of the key data protection principles within the GDPR,[707] and requires organisations to take all reasonable steps to ensure the data they process is not incorrect or misleading.[708] Although this data principle is applicable to AI that processes personal data, the focus is centred on the accuracy of personal data inputted into a system, rather than the system itself and consequential output decisions. In respect of this, future regulation can benefit and complement the GDPR,[709] by providing further clarity and standards.

Accuracy is one of the most important element in AI models, and suggestions to improve accuracy rates in the training phase include increasing the amount of data used, adjusting features and variables, and the combining of different models.[710] Although accuracy in AI systems should be substantially encouraged, several systems are used for predictive methods, and therefore firstly, it would be impossible to measure decisions as accurate or inaccurate at the time they were made, and decisions may be more alike to an opinion, than fact. Also, AI learns by using new

[704] Entilic, ‘Intelligence that cares’ (Corporation Website) <https://www.enlitic.com> accessed 2nd January 2022.
[705] Tom Standage, ‘Automation and Anxiety’ *The Economist* (Special Report, 25th June 2016 Edition); Mary Beth Massat, ‘Artificial Intelligence in Radiology: Hype or Hope?’ [2018] 47(3) *Applied Radiology* 22.
[706] European Union Agency for Fundamental Rights, *Getting the Future Right, Artificial Intelligence and Fundamental Rights* (2020) 35.
[707] GDPR (n 1).
[708] Information Commissioner’s Office, *Guide to the General Data Protection Regulation (GDPR)* (2018, last updated: March 2022) 32.
[709] GDPR (n 1).
[710] Matthew McMullen, ‘How to Improve Accuracy of Machine Learning Model?’ (Medium, 19th August 2019) <https://cogitotech.medium.com/how-to-improve-accuracy-of-machine-learning-model-5ee122727dc1> accessed 22nd September 2022.

data received when in use, by combining it with the previously existing dataset to make future decisions, and consequently, the accuracy of machines may be affected. For these reasons, not only is it imperative that accuracy is monitored at the development stage of AI, but also that obligations exist continuously whilst the technology is in use, to have the ability to measure the accuracy of on-going and past decisions.

The accuracy of AI should also be placed at a high standard once the systems have been developed and are available for public use, to ensure the technology is working as intended. For this to be achieved, decisions made would need to be collected and evaluated regularly for accuracy rates to be produced, also providing the opportunity to alter algorithms in scenarios where inaccuracies have been identified. To encourage AI in achieving an ethical and human-centric status, a proposed new framework could place obligations ensuring the regular review of systems using human judgement, and ensure that clear rules exist to indicate sufficient standards of monitoring. Obligations should also be placed on developers to ensure a high standard of accuracy before such technology is available for deployment. In most scenarios, the developers, and use of that system are independent of one another, and therefore, collaboration from developers to deployers could also benefit the practicalities of implementing such obligations.

It is important to note that the most accurate machines still would not automatically equalise to the most ethical machines. Although accuracy is integral for machines to reduce the likelihood of incorrect outcomes, a high accuracy rate is needed in conjunction with several other principles to ensure the strongest protection to fundamental rights. The Babylon Health App can exemplify this point, which was in use across the UK with partnerships in place with the NHS, but closed after bankruptcy in 2023.[711] The Health App was an inference engine based on ML, and used NLP to offer unparalleled access to healthcare, including treatment advice, assessments, and virtual appointments.[712] Although the app claimed an 82% rate in diagnosis accuracy, which is above the pass-mark for a human qualifying in a

[711] Grace Browne, 'The Fall of Babylon Is a Warning for AI Unicorns' (Wired Blog, 19th September 2023) <https://www.wired.co.uk/article/babylon-health-warning-ai-unicorns> accessed 4th January 2022.
[712] ibid.

parallel role,[713] several safety and data protection concerns were raised, with comparisons being made to Facebook and its use of data to build links and prompt action by consumers.[714] These issues reflect the need for accuracy standards to not be set interdependently, but instead, to be required in conjunction with other principles and requirements.

AI Fairness

To further increase public trust in AI, and for such technology to be accepted, AI systems must be fair. Although 'fairness' in this context has several definitions throughout the literature, an interpretation provided within the EC's Guidelines for Ethical AI can be used as an example.[715] To be fair, AI systems must never lead individuals to be deceived or unjustifiably impair their freedom of choice, and the balance between competing interests and objectives must be respected, ensuring the ability to contest and seek effective redress.[716] 'Fairness' is also included as one of the data protection principles provided by the GDPR,[717] and is interpreted by the ICO as ensuring AI is used in reasonably expected ways, and not used in ways that have unjustified adverse effects, which could lead to unjust discrimination.[718]

For human-centric AI to exist, especially when in public use, it is essential that algorithms are fair, and do not unfairly discriminate, to ensure sufficient safeguards for society. The EU Agency for Fundamental Rights has made their view clear that the principle of non-discrimination, protected under Article 21 of the Charter,[719] should be held at the highest consideration when applying algorithms in everyday life.[720] Borgesius agrees that current non-discrimination principles can be particularly helpful in already prohibiting some types of algorithmic discrimination, but highlights

[713] Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Mobasher Butt, Arnold DoRosario and Saurabh Johri, 'A comparative study of Artificial Intelligence and human doctors for the purpose of triage and diagnosis' [2020] 3 *Front. Artif. Intell.* 543405.

[714] Aliya Ram and Sarah Neville, 'High-profile health app under scrutiny after doctors' complaints' *The Financial Times* (13th July 2018).

[715] High-Level Expert Group on Artificial Intelligence (n 100) 12.

[716] ibid 12.

[717] Alexandra Ebert, 'We Want Fair AI Algorithms – But How To Define Fairness?' (Mostly AI Blog, 6th May 2020) <https://mostly.ai/2020/05/06/we-want-fair-ai-algorithms-but-how-to-define-fairness/> accessed 19th January 2021.

[718] Information Commissioner's Office (n 208) 40.

[719] Charter of Fundamental Rights of the European Union (n 50) Article 21.

[720] European Union Agency for Fundamental Rights, *#BigData: Discrimination in data-supported decision making* (2018, FRA Focus) 1.

the issue of lack of enforcement, and that individuals not being aware of the factors behind decision-making reveal a severe weakness of current protections.[721]

The EP has highlighted the concern towards algorithmic systems producing decisions that not only pose an infringement on individual rights, but also that may give differential treatment to individuals based on their characteristics.[722] This could cause significant issues when algorithms have the power to, for example, offer access to education and employment, or be used by law enforcement.[723] The use of ML by law enforcement has been the subject of discussion in several case-law, including the *Ligue des droits humains* case where the AG has made clear that such systems which produce large numbers of false positives, without the ability to explain how these matches have been made should not be used.[724]

The use of AI by law enforcement was also the focus in *R(Bridges)*[725] in which it was found that the use of FRT imposed an unlawful infringement on Article 8.[726] It was also found that the DPIA did not fulfil statutory requirements,[727] and that the Equality Act had also been undermined, in that not everything reasonable had been done to ensure the software used was free from racial and/or gender bias.[728] This judgment highlights the principle of proportionality, in that, there must be a fair balance struck between individual rights and the public interest, and that for such systems to be used, more effort needs to be made to ensure appropriate policies are in place to assess and protect individual's fundamental rights and freedoms.

Particularly concerning law enforcement, where bias has long been a concern in their practice, it is essential that safeguards are put in place to protect individuals before systems are used and further cases arise in the courts. It can be argued such systems should be prohibited, a view which is supported by several prominent

[721] Borgesius (n 4).
[722] European Parliament, *Resolution on Fundamental Rights Implications of Big Data: Privacy, Data Protection, Non-Discrimination, Security and Law-Enforcement* (2017, 2016/2225(INI)) 19.
[723] ibid 19.
[724] *Ligue des droits humains v Conseil des ministers* (n 702).
[725] *R(Bridges) v CC of South Wales Police* (n 49).
[726] ECHR (n 3) Article 8.
[727] Data Protection Act 2018, s64.
[728] Equality Act 2010, s149(1).

organisations.[729] Law-makers, public bodies and data protection authorities must respect fairness as a concept when using AI, to ensure that any possible measures to minimise issues of discrimination are used.

One complexity that arises in relation to fairness is the number of interpretations throughout the literature, and the difficulties in applying fairness to AI. For fairness to be achieved, the context of the given situation needs to be known, in addition to the possible vulnerabilities that exist in relation to the scenario. AI technology does not have the competency to understand 'fairness' as a concept, which reflects the need for human judgement throughout the training phase, in addition to when systems are in use and analysing new data.

A notable example of unfairness is reflected within a report released by Human Rights Watch, who document how the UK Government's use of a flawed algorithm for Universal Credit led to financial hardship and food insecurity, in addition to affecting individual's mental health.[730] The AI used to assess rates of Universal Credit was based on a means-testing algorithm, which analysed the total income in one complete calendar-month, and then used this information to calculate the amount of Universal Credit to be sent.[731] The algorithm did not take into account how frequently the individual is paid, posing a significant issue to those individuals who are not paid on the same day each month.

Commented on by Human Rights Watch, this design flaw within the means-testing algorithm led to irrational fluctuations and reductions in the amount individuals received,[732] affecting those already most in need. This design flaw was ordered to be fixed by the Court of Appeal in June of 2020,[733] in which the Secretary of State's refusal to put in place a solution to rectify the issue was held to be irrational, under the strict standard of the Wednesbury unreasonableness test.[734] This decision,

[729] EDRi (and 44 others), *Civil Society Calls on the EU to* Prohibit Predictive and Profiling AI Systems in Law Enforcement and Criminal Justice (2022).

[730] Human Rights Watch, *Automated Hardship* (September 2020).

[731] Human Rights Watch, 'UK: Automated Benefits System Failing People in Need' (Blog, 29th September 2020) <https://www.hrw.org/news/2020/09/29/uk-automated-benefits-system-failing-people-need> accessed 19th January 2021.

[732] Human Rights Watch (n 740).

[733] *Secretary of State for Work and Pensions v Danielle Johnson, Claire Woods, Erin Barrett, Katie Stewart* [2020] EWCA Civ 778 [2020] 6 WLUK 270.

[734] *Associated Provincial Picture Houses Ltd v Wednesbury Corporation* (1948) 1 KB 223.

however, did not apply consistently to all who were effected, leaving several individuals without redress, including those paid weekly, fortnightly, or every four weeks.[735] This flawed algorithm threatens the UK's international human rights obligations under Article 22 of the UDHR, which provides the right of everyone to social security, including social insurance.[736] It is also evidently failing those individuals in need, and further pushing individuals into poverty, standards of which do not represent ethical AI.

Several ways have been proposed to aid in rectifying the flaw in the algorithm,[737] however the UK Government remain yet to act. There has been no statement to express whether affected individuals will be compensated for their losses and consequential financial difficulties due to this flaw in the algorithm.[738] This case reflects several issues related to state use of AI and the damaging impact it can have on society. In retrospect, this case highlights the importance of the testing and training phase of AI systems, in which multiple variables should be tested to ensure the minimum risk of error or misunderstanding by the algorithms. In addition, although the algorithm used can be defined as 'unfair', it is still currently in use, continuing to likely affect tens and thousands of people.

Considering this, Human Rights Watch express that the UK Government need to show moral commitment to find solutions for those individuals who are currently being failed by the system.[739] The algorithm in question should have been rectified as soon as possible, and ideally, on a more general basis, an obligation should be placed to ensure flaws are fixed within a specified amount of time. This would not only provide safeguards to individuals subject to decisions, but also give clear standards to organisations. For such flaws to be rectified, firstly they need to be identified, reflecting fairness as another aspect that must be monitored within AI, not only before deployment, but also once systems are in use. Organisations would also be encouraged to employ AI-focused roles to help comply with such obligations, and

[735] Human Rights Watch (n 741).

[736] International Covenant on Economic, Social and Cultural Rights (ICESCR) 1966, Article 22.

[737] Joseph Rowntree Foundation, *Written Evidence to Parliament* (UCW0076, April 2020); Josephine Tucker and Dan Norris, *Rough Justice* (2018, Child Poverty Action Group) 23-25; House of Lords, *Universal Credit isn't working: proposals for reform'* (2nd Report of Session 2019-21, Economic Affairs Committee, published July 2020) 77-85.

[738] Human Rights Watch (n 740).

[739] ibid.

to ensure that individuals have the competency and expertise to rectify such issues promptly.

Several proposals to address the issue of fairness in AI overlap with these ideas, including the need for regular assessments of the representativeness of datasets, ensuring competent humans are involved with assessments, and making algorithms transparent.[740] There are also several schemes in the process of being developed to certify systems as not exhibiting unjustified bias, including the IEEE P7003 Standard, which aims to provide developers with a framework to prevent unintended and unjustified differential outcomes for users.[741] Although a certification process would aid in ensuring algorithmic systems have achieved a certain status before they are deployed, it is imperative that deployers of systems make a consistent effort to check for, and correct any unintended and inappropriate differential outcomes once the system is in use. An open-source toolkit, such as the AI Fairness 360 could be used post-deployment to achieve this, which includes help for deployers to examine, report and mitigate any evidence of bias and discrimination within machines throughout their lifecycle.[742] If a toolkit such as this was standardised, and deployers were legally required to use it, these concerns could be addressed in a much more effective manner.

Another aspect highlighted by the case of *Secretary of State for Work and Pensions v Johnson* is the lack of information made available on the decision-making process used, and the options of redress.[743] Described as unrealistic within the judgment, the Secretary of State had previously suggested that the respondents in this case ask for their salary pay date to be changed, for the algorithm to correctly calculate their payments.[744] It must be ensured that in scenarios where automated decision-making is used, individuals affected should be made aware of its use and workings, and accessible ways to challenge decisions are made available. For organisations to

[740] Rodrigues (n 242); European Parliament (n 732); Bettina Berendt and Soren Preibusch, 'Toward Accountable Discrimination-Aware Data Mining: The Important of Keeping the Human in the Loop-and Under the Looking Glass' [2017] 5(1) *Big Data* 135.

[741] Rodrigues (n 242); IEEE, 'Algorithmic Bias Considerations' (2017, P7003 Project) <https://standards.ieee.org/ieee/7003/6980/> accessed 10th January 2022.

[742] IBM, 'AI Fairness 360' (2018, Corporation Website) <https://aif360.res.ibm.com/> accessed 5th January 2022.

[743] *Secretary of State for Work and Pensions v Danielle Johnson, Claire Woods, Erin Barrett, Katie Stewart* (n 743) [97, 99].

[744] ibid [96].

produce sufficient information on AI systems, firstly, individuals within the organisation itself would need expertise and understanding, and secondly, the AI systems used would need a sufficient degree of transparency. Organisations having transparency themselves with the technology they use, and how it benefits them could not only improve consumer relations, but also encourage and promote public trust in AI.

In addition to those who deploy AI, a general national effort should be made to raise awareness, and educate the public on AI technology, including its progression, the benefits and drawbacks, and the potential impact to society. For education on AI to strengthen public trust, legislative measures need to be put in place to provide assurance and security to the public. This would ensure that even if the drawbacks of AI were explored extensively through an educational campaign, the legislative measures and safeguards that exist can also be made available, to reflect the efforts made to minimise the risks. Such measures should also provide clear guidance on redress procedures, and obligations to ensure options of redress are accessible, and dealt with by human judgement and in a timely manner. This again reflects the importance for organisations to have AI competent employees, not only to ensure errors are dealt with quickly, but also to be able to discuss issues with individuals, and following given guidance, can offer an adequate solution.

AI Bias

To promote and encourage the use of ethical AI, mitigating bias in algorithms is crucial, not only to enable all of society to benefit from AI, but also to provide sufficient safeguards to the public. AI bias exists when an algorithm produces prejudiced results, due to flawed assumptions within the decision-making process,[745] and stems from two major variables. AI bias can occur due to non-diverse datasets inputted during the training and development process, or in scenarios where AI itself classifies the wrong data for the wrong purposes, consequently, producing bias.[746] Considering the rapid progression of use by the state and major corporations, it is fundamental that sufficient safeguards are in place to protect the whole of society.

[745] Schwartz (n 327); Dastin (n 331).

[746] Dastin (n 331).

Evidence of bias in AI is widespread through reports and the literature,[747] and for ethical AI to exist, bias must be mitigated, and tested for during development phases, in addition to when AI is in use, and throughout its lifecycle.

Bias existing within AI can cause a devastating impact to society and individuals, evidently seen in the most notable example of the 'COMPAS' system,[748] used by the US courts to predict the likelihood of defendants becoming a reoffender. The sentencing algorithm in question predicted twice as many false positives for offenders of black ethnicity, in comparison to those of white ethnicity,[749] posing a significant interference with individuals human rights protections.[750] In addition to this, a study in 2019 found that an algorithm used for over 200 million people in the US healthcare system, to predict patients who would likely benefit from extra medical care, favoured individuals of white ethnicity over those of black ethnicity. [751] The algorithm itself did not take race into account, however, did consider individuals previous health care spending, reinforcing pre-existing inequities in the US healthcare system, in which less money is reportedly spent on medical care for patients of black ethnicity, in comparison to those of white ethnicity, who have the same level of need.[752]

Due to this disparity, the algorithm scored white patients as higher risk and more in need of extra healthcare in comparison to black patients, perpetuating the bias that has already long existed.[753] However, since this bias was identified, with aid from researchers, developers were able to reduce the levels of bias by 80% through changing the variables used.[754] This reflects not only the importance in sufficiently

[747] AI Now, *Disability, Bias and AI* (2019); W. Nicholson, 'Medical AI and Contextual Bias' [2019] 33 *Harv. J.L. & Tech.* 66; Centre for Data Ethics and Innovation, *Review into bias in algorithmic decision-making* (2020); Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, 'Machine Bias' (ProRepublica, 23rd May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 11th January 2021.

[748] Angwin, Larson, Mattu and Kirchner (n 757); Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, 'How we Analyzed the COMPAS Recidivism Algorithm' (23rd May 2016) <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> accessed 12 January 2021.

[749] Angwin, Larson, Mattu and Kirchner (n 757).

[750] ECHR (n 3) Article 14; Equality Act 2010, s4.

[751] Ziad Obermeyer, Brian Powers, Christine Vogeli and Sendhil Mullainathan, 'Dissecting racial bias in an algorithm used to manage the health of populations' [2019] 366(6464) *Science* 447.

[752] ibid.

[753] ibid.

[754] Terence Shin, 'Real-life Examples of Discriminating Artificial Intelligence' (Towards Data Science Blog, 4th June 2020) < https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070> accessed 12th January 2021.

monitoring systems in order to identify bias, but also the opportunity provided to rectify issues. Bias have not only been evidenced in the US, but also in the UK, where the Court of Appeal found that law enforcement had failed to recognise the risk of gender and racial discrimination through the deployment of FRT,[755] an obligation imposed by the Equality Act.[756]

AI expert Sharkey emphasises that automated decision-making algorithms are currently propagating gender and race discrimination through global communities.[757] In an interview to the guardian, Sharkey calls for moratoriums to be imposed on all 'life changing' decision-making algorithms in Britain, arguing that algorithms are so 'infected with biases' that systems used should not be trusted, and are not fair.[758] To solve this issue, Sharkey recommends pharmaceutical-style testing, in which systems would need to be tested on millions of people in order to reach a sufficient point that reflects no evidence of a major inbuilt bias.[759]

The Committee on Standards in Public Life also supports the need for further guidance and regulation, in the cases of transparency and data bias in particular.[760] In addition to this, the commendable review into bias in algorithmic decision-making was released by the Centre for Data Ethics and Innovation in November 2020, and highlighted the urgent need for improvement worldwide in addressing algorithms to promote and encourage ethical AI, not undermine it.[761] In respect of this, they suggest that regulation can aid in addressing AI bias by setting minimum standards, clear guidance for organisations to meet their obligations, and proper enforcement to ensure minimum standards are met.[762]

For bias to be identified in systems, firstly, algorithms must have enough data within its data set for a bias to be reflected, and secondly, it is essential that systems are

[755] *R(Bridges) v CC of South Wales Police* (n 49).
[756] Equality Act 2010, s149(1).
[757] Noel Sharkey, 'The impact of gender and race bias in AI' (Humanitarian Law & Policy Blog, 28th August 2018) <https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/> accessed 12th January 2021.
[758] ibid; Henry McDonald, 'AI expert calls for end to UK use of 'racially biased' algorithms' *The Guardian* (12th December 2019).
[759] ibid.
[760] The Committee on Standards in Public Life, *Artificial Intelligence and Public Standards* (February 2020, Independent Report) 7.
[761] Centre for Data Ethics and Innovation (n 757) 3.
[762] ibid 11.

thoroughly tested using human judgement before they are deployed, in addition to sufficient monitoring throughout their lifecycle. Requirements should also be put in place for developers of AI technology to ensure diversity within data sets,[763] to mitigate bias in the training phase of systems and provide further safeguards to minimise risk. A clear and efficient effort to mitigate bias in machines is important for developers in recognising and tackling the challenges of achieving ethical AI, as well as building and encouraging public confidence. Ofcom are also in support of this view and suggest testing regimes and auditing as possible solutions to check for bias, through calibrating datasets to understand how representative they are of society.[764]

Bias can form and lead to discrimination of several different characteristics protected under the Equality Act.[765] This makes it essential that algorithms are checked for any evidence of bias through the testing and monitoring stages, to ensure these protections are upheld. Another aspect the thesis advocates for is the need for organisational enforcement, in which for example, an authority, could subject organisations, whether random or targeted, to an independent review of their algorithms in use, to not only ensure compliance with other proposed requirements, but also to inspect algorithms as a further safeguard to individuals and their rights, and in encouragement and promotion of ethical AI. This independent review could also be made available to the public, and could provide pressure on organisations to ensure systems are working to a sufficient standard to protect not only their reputation, but society as a whole.

3.2.2 Transparency, Explainability and Interpretability in AI systems

AI Transparency

For the **accuracy**, **fairness** and **bias** explored above be identified in AI, and for sufficient options to challenge decisions be made available, it is essential that AI technology is subject to mandatory transparency requirements. Transparency is listed as a key requirement within the EC's Ethics Guidelines for Trustworthy AI,[766]

[763] Shin (n 764).
[764] Cambridge Consultants (n 4).
[765] Equality Act 2010, s4.
[766] High-Level Expert Group on Artificial Intelligence (n 100).

and includes the concepts of traceability, explainability and communication.[767] Traceability refers to the documentation of the data sets and processes used by AI to arrive at a decision, also applying to past decisions, and aids in enabling reasons as to why a decision in question may have been flawed, and consequently, helps to prevent future errors.[768]

Explainability, discussed further in the next section, relates to the ability to explain the technical processes of AI, and that it be traced and understood by a human.[769] The concept of communication refers to the importance of citizens being made aware of AI technology and when it is used, expressing that individuals have the right to be informed when interacting with AI, and that an option be made available to opt for human interaction.[770] To provide more clarity, the thesis' recommendations for a proposed new framework for AI advocates for a minimum standard for transparency in AI systems, not only to make redress options effective and practical, but also to ensure organisations can comply with the proposed testing and monitoring requirements.

As noted within the literature, full transparency in machines would result in a counterproductive approach and would lead to an information overload, or make it possible for individuals to game systems.[771] Also, the EDPB and EDPS comment that there may be scenarios where it is arguable that little to no transparency can be given to the public due to reasons of secrecy,[772] in which it is still imperative that safeguards be put in place to ensure such systems do not unlawfully infringe on fundamental rights. The EDPB and EDPS suggest that in these cases, systems should be registered and provide transparency to a competent supervisory authority,[773] however, more guidance may be needed in terms of which supervisory body should govern this burden. For transparency to be effective, it must be ensured that the decision-making processes and data sets used are made available, and that information can also be produced on individual decisions, to aid in redress options.

[767] ibid 18.
[768] ibid 18.
[769] ibid 18.
[770] ibid 18.
[771] Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan and Wallach (n 349).
[772] European Data Protection Board, *EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act)* (2021) 20.
[773] ibid 20.

In support of this, Bickerstaff expresses that a legal framework needs to be developed to include clear transparency requirements,[774] which, commented by the EDPB, is a very challenging goal.[775]

Emphasised further in the review produced by the Centre for Data Ethics and Innovation into algorithmic decision-making, it must be ensured that decisions can be scrutinised, explained and challenged, in order to prevent legislative measures from losing their effectiveness.[776] In addition to this, organisations must make it clear when AI is in use, and could be placed under an obligation to provide accessible information on the technology they use and for what purposes, including the benefits it provides to the organisation in question, as well as the redress options for individuals. This approach has also been considered in the new ranking guidelines under the P2B Regulation,[777] which although are not legally binding, aim to promote fairness and transparency for online intermediation services.

For legislation to be effective long-term, it must also account for future issues in which may not currently be known, or have the flexibility for the scope to be amended in the future. In respect of this, Engler considers that AI is still a new and modern technology, and that setting a sufficient standard for transparency may aid for reasons not apparent today.[778]

AI Explainability

Explainability and interpretability are widely recognised as neighbouring concepts to transparency, and are equally important in promoting and encouraging the progression of ethical AI. Whilst standards of transparency would give information in regard to the decision-making processes used, and the ways in which systems were

[774] Roger Bickerstaff and Aditya Mohan, ‘A ‘Light Touch’ Regulatory Framework for AI – Transparency at the Heart of AI Regulation’ (Digital Business Law Lexology Blog, ND) <https://www.lexology.com/library/detail.aspx?g=33697021-f272-4d51-ad0f-2e1d87e0857d> accessed 14th January 2021.

[775] European Data Protection Board (n 782) 20.

[776] Centre for Data Ethics and Innovation (n 166) 6.

[777] Regulation (EU) 19/1150 on promoting fairness and transparency for business users of online intermediation services [2019] OJ L 186; Digibyte, ‘European Commission publishes ranking guidelines under the P2B Regulation to increase transparency of online search results’ (December 2020, EU Website) <https://digital-strategy.ec.europa.eu/en/news/european-commission-publishes-ranking-guidelines-under-p2b-regulation-increase-transparency-online> accessed 6th January 2022.

[778] Alex Engler, *The Case for AI Transparency Requirements* (The Brookings Institution’s Artificial Intelligence and Emerging Technology Initiative, January 2020).

trained, the concept of explainability is stated to be the ability for AI systems to explain the reasons behind a certain decision.[779] From a technical perspective, the term ‘explainability’ refers to the ability of an AI system to present its processes and decision-making logic in an understandable way to humans.[780] In regulatory terms, explainability maps onto transparency closely, where the goal is not only to interpret technical outputs, but also to provide accountability in decision-making processes. Explainability is also not considered within the EU’s White Paper for AI, however, the ability for AI technology to be able to provide explanations on decisions reached is a key component to ethical AI.[781]

Explainability is essential for options of redress to be effective, and to allow unpredictable decisions be understood and addressed.[782] However, an obligation to comply with standards of explainability could pose a limit, on those machines noted in the literature to be too complex to have the ability to provide explanations for their decisions.[783] Regardless, to promote and encourage human-centric ethical AI, it may be essential for such limitations to exist, at least until further effective future solutions become available.

The most debated example of a right to explanation can be seen within the GDPR, in which Articles 13-15 provides an obligation on data controllers to provide data subjects with meaningful information about the logic involved of decision-making, as well as the significance and the envisaged consequences of such processing for the data subject in scenarios where automated decision-making is used.[784] Although a more explicit right to explanation is stated within the non-binding recitals of the GDPR,[785] the terminology used remains vague and lacks clarity.[786] It is essential that a future framework for AI sets a clear standard for explainability, and that ways to achieve such a standard should be made publicly available. In respect of this, and in

[779] Wachter, Mittelstadt and Russell (n 523).
[780] Fatima Hussain, Rasheed Hussain and Ekram Hossain, ‘Explainable Artificial Intelligence (XAI): An Engineering Perspective’ (January 2021) <https://www.semanticscholar.org/reader/1f0d09386ee7685c4a8953aed81adbd4055763c1> accessed 4th September 2024.
[781] European Parliament, *Understanding algorithmic decision-making: Opportunities and challenges* (2019) 3.
[782] ibid 4.
[783] Information Commissioner’s Office (n 518).
[784] GDPR (n 1) Article 13(2f), Article 14(2g) and Article 15(1h).
[785] GDPR (n 1) Recital 71.
[786] Edwards and Veale (n 179).

support of the suggestion given by Bickerstaff and Mohan, a public transparency and explainability database or register should be established by legislation, not only for multi-organisational use, but also so registered approaches can be scrutinised,[787] and consequently improved for future use. A public database for approaches to AI technology could also benefit other aspects, including methods to mitigate bias, and to improve fairness and accuracy.

Explanations can be produced and given in several ways, and different approaches may be more appropriate in various scenarios. The literature has suggested a lack of consensus towards defining a ‘general explanation’ and instead, implies that an effective explanation would differ depending on the algorithm used, and the context of its use.[788] Irrespective of this, a proposed new legislation could aid in producing a minimum standard to ensure organisational compliance with structured options of redress, in which a higher standard could be held for those decisions that have more substantial effects on individuals. For individuals to be able to understand and challenge decisions made, the explainability concept should also be mandatory for AI technology in use, to ensure any information given or made public can be understood by a reasonable layman.

Enforcement of standards such as this could be achieved in combination with transparency requirements, a view which is agreed with by the EDPB, who state that explainability requirements should also provide for additional transparency.[789] In this regard, obligations could be placed on organisations to raise awareness to individuals that are subject to AI decisions, and that clear options to understand and challenge decisions are made publicly accessible. In this respect, a public educational campaign on AI could also raise awareness that such information is readily available from organisations, and that sufficient redress options exist. Clear rules and guidance within a proposed new framework could not only aid in ensuring organisational compliance, but also raise public awareness and consequently improve public trust of AI, by ensuring effective and timely redress options are available.

[787] Bickerstaff and Mohan (n 784).
[788] Alejandro Barredo Arrieta and Francisco Herrera, ‘Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges towards Responsible AI’ [2020] 58 *Information Fusion* 82.
[789] European Data Protection Board (n 782) 18.

AI Interpretability

The concept of interpretability can be understood as a passive feature to AI, and refers to the extent an algorithm can be understood.[790] A more detailed explanation of interpretability includes the degree a human can estimate a system's prediction, the extent a human can understand and follow the decision-making process that led to a prediction, and the extent to which a human can detect an error within a system.[791] AI interpretability is understood as an umbrella term which includes the concepts of transparency and explainability, but centres on the ability to predict the outcome of machines.[792] Arguably, higher levels of interpretability in AI will amount to an increase in public trust in systems, and this may increase further the more humans are able to demonstrate an understanding. Interpretability is not seen as necessary to impose on systems that cause minimal effects in comparison to those that cause more substantial effects.[793]

A high level of interpretability in machines would allow for a more in-depth comparison between two competing systems, including the ability to justify why one model may be better than the other, and identification of aspects that are in need of improvement within machines.[794] The concept of interpretability would need to be considered within a future framework for AI, in order for sufficient testing and monitoring requirements to work in practice, and for such measures to achieve the aim intended, to identify errors and flaws in machines, and for them to be rectified. Jackson expresses that the inclusion of these concepts is paramount to ensuring justice, in particular when AI deployment is widespread across multiple industries.[795]

[790] Gilpin, Yuan, Bajwa, Specter and Kagal (n 345).
[791] Aji Thampi, 'Interpretable AI or How I Learned to Stop Worrying and Trust AI' (Towards Data Science, 5th March 2019) <https://towardsdatascience.com/interpretable-ai-or-how-i-learned-to-stop-worrying-and-trust-ai-e61f9e8ee2c2> accessed 15th January 2021.
[792] ibid.
[793] Jonathan Johnson, 'Interpretability vs Explainability: The Black Box of Machine Learning' (BMC Blog, 16th July 2020) < https://www.bmc.com/blogs/machine-learning-interpretability-vs-explainability/> accessed 15th January 2021.
[794] ibid.
[795] Brandon Jackson, 'Artificial Intelligence and the Fog of Innovation: A Deep-Dive on Governance and the Liability of Autonomous Systems' [2019] 35(4) *Santa Clara High Technology Law Journal* 35.

Interpretability standards may also aid in allocating liability, with reference to establishing a 'reasonable computer' standard proposed by Abbott.[796] For these reasons, demands for interpretability requirements for AI are arguably justified, however, focus needs to be made on the level of interpretability in systems. Yampolskiy acknowledges that the inherent nature of unpredictability within AI technology will forever ensure 100% completely safe AI cannot be achieved, however, safer AI should always be strived for.[797]

Demands made for complete interpretability in AI prior to related systems being available for use would be extremely problematic and unrealistic, as this would suggest that a human counterpart could predict 100% of AI decisions, and therefore be capable of making the same decisions in the same timescale. Across the literature, the main benefit of AI is stated to be the ability to make decisions at an increased speed and efficiency in comparison to humans, and with theories of future AI systems having stronger and smarter capabilities than humans,[798] demanding complete interpretability pre-deployment would be inappropriate.

Another very well-known benefit of AI is the ability for it to do the tedious and laborious work of humans in a much more efficient manner, arguably reducing the need for humans to be in these positions. In reflection of this, demanding those who develop AI to have a full understanding of the systems and their predictabilities would reduce this benefit of AI systems, and could question whether AI should even be used if a human needs to oversee, understand and be able to make the exact same decisions using the exact same processes.

A high level of interpretability could possibly be achieved, however would pose a substantial restriction and limitation on the novel concepts within AI technology, mainly in reference to its ML capabilities. This would also cause an obvious and significant effect to the AI industry, and consequently stifle innovation, deeming several existing systems pointless. These views strike the importance of developing

[796] Abbott (n 402).
[797] Roman Yampolskiy, 'Unpredictability of AI' (2019) <https://arxiv.org/ftp/arxiv/papers/1905/1905.13053.pdf> accessed 15th April 2021.
[798] Darrell West and John Allen, *How Artificial Intelligence is Transforming the World* (Brookings EDU Report, 2018).

a balance between demands for interpretability and ensuring the encouragement and promotion of the progression of AI. Jackson also reflects on the balance needed in consideration on interpretability, commenting that there is a fine line between regulatory actions that will achieve in infusing AI with democratic values of fairness, and those that will amount to barriers to innovation.[799]

3.2.3 A Proposed New Framework for AI: The Recommended Ethical Safeguards

For such systems to contain the ethical principles above, effective enforcement is imperative. A future framework for ethical AI would not only provide a legal basis for enforcement, but also strengthen and clarify safeguards for both organisations and those affected through the use of AI, to ensure sufficient protection to individuals and their rights. This section, with respect to the principles discussed above, aims to answer RQ1 by clearly stating how a proposed new framework for AI could address the challenges posed to ethics and human rights. The recommendations include pre-emptive measures that should be achieved before AI technology is deployed, including requirements to ensure the testing of AI meets a sufficient standard, for a FRIA to be completed, and rules regarding the datasets used for training purposes.

In the context of ensuring enforcement, the current attempts by the ICO and the GDPR should be noted to highlight where improvement is needed. Recital 148 of the GDPR provides guidance to strengthen enforcement, by stating that penalties and administrative fines should be imposed for any breaches of the regulation by the supervisory authority,[800] which in the case of the UK, is the ICO. One of the ICO's main functions is to ensure national adherence to the GDPR, a principle stressed within the *Schrems II* case.[801] In this context, the BILETA response highlights the concern that despite the above, there has been evidence that the ICO has overlooked several serious matters, due to most of its resources being devoted elsewhere to enforcement obligations.[802] it should be ensured that the relevant

[799] Jackson (n 804).
[800] GDPR (n 1) Recital 148.
[801] Case C-311/18 *Data Protection Commissioner v Facebook Ireland Limited and Maximilliam Schrems* [2020] OJ C 249 [108].
[802] Felipe Romero Moreno, Edina Harbinja, Henry Pearce, Karen Mccullagh, Gavin Sutter, Subhajit Basu, Orla Lysnkey and Aysem Diker Vanberg, *BILETA Response to UK Government Consultation: Data a new direction* (2021) 51.

authority has enough resources to enforce obligations, which may call for an entirely separate authority to focus on such obligations.

These rules would ensure high rates of **accuracy**, **fairness**, and **non-discrimination** in machines before deployment, and aim to minimise the risk of errors or flaws in algorithms before available for public use, therefore encourage and promote the progression of ethical AI. The recommendations proposed also suggest post-deployment measures, including strict and enforceable monitoring obligations, to ensure AI remains working at the required standards of **accuracy**, **fairness**, and **non-discrimination** throughout their lifecycles. Strict requirements to enforce meaningful monitoring of systems also allows the opportunity to identify and rectify any errors that have developed since deployment, to ensure AI technology remain working in the ways intended.

As noted by Algorithm Watch, fundamental rights will only be sufficiently protected once relevant legal provisions can be effectively enforced.[803] An authority should be established within new legislation, and granted powers to ensure organisational compliance, and consequently the protection of citizens and their fundamental rights. In the EU Agency for Fundamental Rights report on AI, it is acknowledged that it is not clear to public administration or to organisations, who is responsible for checking and overseeing the use of AI.[804]

Although the ICO have been directly involved in clarifying AI guidance, particularly in reference to the GDPR, the report highlights that the data protection authority are under-resourced for the additional task, due to the lack of relevant AI-related expertise, and the heavy workload and overstretched budget currently faced.[805] The recent AI Act advocates for a European Artificial Intelligence Board (EUAIB) to be established,[806] comprising of representatives from MS and the EC, to facilitate the implementation of regulatory rules, offer advice and expertise, and share best practices, a recommendation which this thesis strongly supports, although national

[803] Mackenzie Nelson and Friederike Reinhold, ‘The DSA Proposal is a good start. Now policymakers must ensure that it has teeth’ (Algorithm Watch, 16th December 2020) <https://algorithmwatch.org/en/dsa-response/> accessed 17th January 2021.
[804] European Union Agency for Fundamental Rights (n 716) 95.
[805] ibid 95; Brave, *DPA Report* (2020) 6-8, 11-12.
[806] Provisional Agreement for the AI Act (n 103) Article 56.

authorities would also be needed to guide enforcement. Of course, the AI Act will not directly apply to England and Wales, but inspiration should be taken from the EU on this point.

It is essential that when new legislation for AI is introduced in England and Wales, a body with the expertise and ability to have a sole focus on AI and its challenges must also be established. Within the thesis' recommendations for a proposed framework for AI, a national 'AI Human Oversight Board' is suggested, with the aim of not only ensuring compliance to obligations, but also providing clarity and guidance for best practices to organisations and developers, to encourage and promote a fundamental rights and ethical approach to the use of AI. The proposed board is also held responsible for guidance on redress matters and ensuring organisational compliance. This thesis echoes the concern highlighted by Leprince-Ringuet that the current UK Government is raising the idea of removing the legal provision which enables individuals to challenge a decision made solely by automated decision-making and to request human review of decisions.[807] Opposing this idea, and particularly in reference to the issues highlighted in the BILETA response regarding the ICO and enforcement,[808] the proposed suggestions aim to work collaboratively to safeguard the protection of fundamental rights, through providing clear rules, and enforcing compliance to promote the progression of ethical AI, complementing the ICO.

Pre-deployment Measures for AI technology: Sufficient Testing and Training

For ethical AI that respects individuals' fundamental rights to exist, it must be ensured that legislative measures address the training and testing of AI technology before deployment, to ensure that systems meet a clear ethical standard. The following measures suggested aim to address the human right and ethical challenges posed by AI systems, to offer clear rules and guidance to developers and organisations, and provide sufficient protections to individuals and their rights. As noted, AI technology makes use of large datasets in the training stages of their

[807] Daphne Leprince-Ringuet, 'Even computer experts think ending human oversight of AI is a very bad idea' (14th October 2021) <https://www.zdnet.com/article/even-computer-experts-think-ending-human-oversight-of-ai-is-a-very-bad-idea/> accessed 19th June 2022.

[808] Romero Moreno, Harbinja, Pearce, Mccullagh, Sutter, Basu, Lysnkey and Diker Vanberg (n 811) 51.

development.[809] These datasets used, as discussed above, have the risk of being responsible for development and amplification of bias.[810] For this reason, an obligation should be placed on developers of AI technology to ensure diversity of data within training data sets, which at the very least, reflects the diversity of potential users of the technology. In practice, this may be difficult to achieve, as there is not a standard definition or process which entails how diversity in datasets really could be achieved, and as noted in the literature, it may be possible for datasets to be diverse without being inclusive.[811]

Data should be assessed before being used within training datasets by developers, to ensure no pre-existing evidence of bias or inaccuracies. To encourage developers to comply with such obligations, a certification procedure could be introduced to recognise efforts in developing diverse and ethical AI and be consequently used by developers to attract consumers. For evidence of bias be identified in a training scenario, the size of the dataset needs to be large enough to be able to reflect any possible bias. In consideration of this, the thesis advocates for use of large datasets in the training phase of systems, whilst also assuring sufficient related data protection security, to ensure such data is not misused.

The use of large datasets would also put developers in an advantageous position to comply with the following pre-deployment recommendations included within the suggestions below. This requirement, however, would contrast with the GDPR’s data minimisation principle, which states that use of personal data should be limited to what is necessary in relation to the purposes for which they are processed.[812] It could be argued that large datasets are indeed necessary in AI systems for the purpose of mitigating bias, but clarification would be needed to ensure this principle would not be breached.

Currently under the GDPR, data controllers must complete a DPIA prior to processing personal data for high risk purposes.[813] In support of the suggestion

[809] European Parliament (n 247) 1.
[810] ibid 1.
[811] Michael Kraus, Brittany Torrez and LaStarr Hollie, ‘How Narratives of Racial Progress Create Barriers to Diversity, Equity, and Inclusion in Organizations’ [2022] 43 *Current Opinion in Psychology* 108.
[812] GDPR (n 1) Article 5(c).
[813] ibid Article 35.

made by the EU Agency for Fundamental Rights and throughout the literature, the thesis advocates for an obligation on developers and organisations to also carry out a FRIA prior to processing.[814] Contrasting the high risk requirement for a DPIA, the obligation for a FRIA should be introduced for all systems which process personal data, to account for those systems which are not primarily or obviously a high risk.

FRIAs also provide the opportunity to address matters related to fundamental rights in addition to, and to matters not included within a DPIA. As suggested by Latonero, sector-specific toolkits could also be made available to combat relevant industry needs and specific issues.[815] This extra focus within FRIAs would also increase awareness to developers of the potential risks AI poses to fundamental rights, and guidance should be made publicly available and accessible on how to mitigate such risks.

Alike to the Algorithmic Impact Assessment tool[816] produced by the Canadian Government under the Canadian Directive on Automated Decision-Making,[817] an online assessment could be established for companies with extensive questions that relate to fundamental rights concerns.[818] The Canadian system produces a final impact score that is made publicly available on the Government's website. Such a system could also be implemented through future legislation, where a minimum score must be achieved before systems are made publicly available. For systems that are more likely to cause potential interference with fundamental rights, a higher score should be required before deployment, to ensure extra and strengthened safeguards to individuals. A similar approach with differing minimum standards dependent on the risks related to AI systems should also be adopted for testing and monitoring requirements.

[814] European Union Agency for Fundamental Rights (n 716) 87; Commissioner for Human Rights, *Unboxing Artificial Intelligence: 10 steps to protect Human Rights – Recommendation* (Council of Europe, 2019); Heleen Janssen, 'An approach for a fundamental rights impact assessment to automated decision-making' [2020] 10(1) *International Data Privacy Law* 76; Mantelero (n 552); Edwards and Veale (n 179); Access Now (n 140).

[815] Mark Latonero, *Governing Artificial Intelligence: Upholding Human Rights & Dignity* (2018, Data & Society) 2.

[816] Government of Canada, 'Algorithmic Impact Assessment' (Canadian Government Website, Algorithm Impact Tool, 2019) <https://canada-ca.github.io/aia-eia-js/> accessed 22nd January 2021.

[817] Canadian Directive on Automated Decision-Making 2019, Article 6 and Appendix C.

[818] European Union Agency for Fundamental Rights (n 716) 90.

As noted, this approach differs from the one taken within the GDPR,[819] and differs from the approach by the EU, which focuses predominantly on high risk systems. For ethical AI to be achieved, ethical standards, practices and obligations need to be placed on all AI systems, to provide safeguards to prevent unintended or unforeseeable consequences which could be produced. Although it is imperative to distinguish different risk levels to proportionately regulate AI, general requirements would aid those systems in the 'grey area', where it is not clear whether the high risk category applies, such as online advertising.[820] This would not only reduce the pressure on deciding the classification of systems, but also avoid those not included left subject to no safeguards at all, making it clear that all AI systems should be held to a high standard regardless of their risk level.

The thesis advocates for the introduction of clear requirements for the testing of AI, including that human judgement should always be present in the review of testing results. To address the challenges posed to human rights, testing should include a review of **accuracy**, **fairness**, and **non-discrimination**, and be repetitive, thorough, and adequately reviewed before the related AI system is deployed. The testing of systems should include a review of a range of changing variables,[821] and a clear standard should be met before systems are made available for use. For such testing requirements to work sufficiently in practice, **transparency**, **explainability** and **interpretability** must be foundational concepts through the development, training, and testing phases of AI deployment. Even if the testing of systems is effective before deployment, there is still a need to ensure this continues once in use.

Post Deployment Measures for AI Technology: Sufficient Monitoring and Organisational Expertise

To ensure AI systems continue to work at the high standard achieved before deployment, a future framework for AI must set clear rules for machines once they are in use, to further promote and encourage a future of ethical AI. To ensure the rights provided within the GDPR in AI systems are upheld, the ICO express that data controllers are required to retain records of all AI decisions made, as part of the

[819] ibid 90.
[820] ibid 90.
[821] ibid 90.

accountability and documentation obligations, including where individuals have requested human intervention, expressed any views, contested the decision, and the results of that challenge.[822] The EDPS also make it clear that any monitoring requirements need to be made in line with the proportionality requirement, to ensure such obligations strike a balance between protecting rights and the public interest.[823]

The ICO adds that this data should be monitored and analysed, to ensure systems are amended accordingly when necessary, to reduce any similar errors occurring in the future.[824] Through analysis of data concerning individuals exercising their rights provided by the GDPR, if any grave or frequent mistakes are identified, data controllers must take immediate steps to understand and rectify the issues.[825] However, with minimal enforcement, and a lack of specified timely obligations about rectifying machines, it can be questioned whether such guidance, and rules, are sufficient for the future of AI.

The case of *Barbulescu*[826] highlights a possible issue when imposing monitoring requirements, in which it needs to be ensured that where an employee is subject to monitoring, measures need to be accompanied by adequate and sufficient safeguards against abuse. In *Rotaru*, it was expressed that adequate and effective safeguards are imperative, as a system of secret surveillance could have the power to undermine, or even destroy democracy.[827] The courts have made it clear that any possible infringement to privacy (or other qualified rights) needs to be proportionate to the aim pursued, and that the subject needs to be protected against arbitrariness,[828] either from prior review by the courts or independent authorities.[829] This can be applied to the monitoring of systems, in which it needs to be ensured that the process is proportionate, and that safeguards need to be in place to protect from abuse. It also needs to be ensured that any infringement of human rights through monitoring is in accordance with the law, and hence respecting the rule of

[822] Information Commissioner's Office (n 718) 221-222.
[823] European Data Protection Supervisor, *Orientations from the EDPS. Reactions of EU Institutions as Employers to the COVID-19 Crisis* (2020) 5.
[824] The Information Commissioner's Office (n 718) 172.
[825] The Information Commissioner's Office (n 718) 34.
[826] *Barbulescu v Romania* App no 61496/08 (ECtHR, 5 September 2017) ECHR 754.
[827] *Rotaru v Romania* (n 461) [59].
[828] *Barbulescu v Romania* (n 835) [121].
[829] Joined Cases C-203/15 and C-698/15 *Tele2 Sverige AB v Post-och telestyrelsenk* [2016] All ER (D) 107 (Dec) and *Secretary of State for the Home Department v Tom Watson* [2016] All ER (D) 107 [123].

law,[830] re-emphasising the need for legislation to set standards, so that this can be upheld.

The thesis supports obligations being placed on those who develop, and deploy systems to monitor AI technology thoroughly and regularly, and that all relevant reviews of the monitoring results should be analysed using human judgement, to provide the opportunity to identify any inaccuracies, or evidence of unfairness or bias in algorithms. The proposed monitoring standards would need to be subject to approval from an independent authority, to ensure that individual rights are respected, whilst also ensuring that systems are working as intended. This improves on the guidance from the ICO, which suggests errors should be identified by analysis of the number of individuals who exercise their rights,[831] and instead adds a pre-emptive feature to potentially improve machines before a flaw or error affects an individual.

A general product monitoring duty on the part of producers does exist within the code of practice in product safety recalls, which expresses that mechanisms to monitor the safety of products should be established.[832] In reflection of this, a future framework could impose an enforceable and direct duty for developers to monitor systems, described as paramount importance for future regulation in light of the characteristics of AI.[833]

For such monitoring requirements to be complied with by developers, the proposed recommendations advocate for organisations to establish AI-specific roles and teams who have the competency, and expertise to work with and establish a relationship with developers to analyse machines sufficiently.[834] The proposed monitoring requirements include thorough and regular testing concerning **accuracy**, **fairness** and **non-discrimination**, to provide opportunities for flaws to be identified, and require results of audits to be collated. For these requirements to work in practice, it is imperative that a future framework for AI is built upon the fundamental concepts of

[830] *Klass and others v Germany.* App No 5029/71 (ECtHR, 6 September 1978) 2 EHRR 214 [55].
[831] The Information Commissioner's Office (n 831) 104.
[832] Office for Product Safety & Standards, *Code of Practice on consumer product safety related recalls and other corrective actions* (PAS 7100:2018) Part 1.
[833] European Commission (n 67) 7.
[834] European Union Agency for Fundamental Rights (n 716) 90.

transparency, **explainability** and **interpretability**, and sets clear and adequate standards to be met, which can be understood and adhered to by those working with the technology.

The recommendations for a future framework for AI intend to go further than the guidance provided by the ICO in reference to the GDPR,[835] by suggesting an obligation on developers to input the related results on an established system, and that they clearly state whether any action is intended to be taken to rectify or improve the related algorithm in question. This system should also be made available to the relevant authorities, so compliance can be documented and accessed, when necessary, by either party. Any evidence of inaccuracies, unfairness or bias identified within the reviews should also be flagged within the proposed system and rectified in a proposed timeframe set by the proposed new authority, for example, the 'AI Human Oversight Board', and dealt with on a case-by-case basis. The introduction of a new focused authority to aid in these matters would ensure developer compliance with such requirements through enforcement power and resources, and consequently would guarantee the purpose of the requirements, to encourage and promote the progression of ethical AI.

These suggestions not only allow errors or flaws in AI technology to be identified, but ensure related flaws are dealt with efficiently, and promptly. The measures proposed ensure that the challenges to ethics and human rights brought by AI are adequately addressed, and provide further security and protection to individuals and their fundamental rights. These post-deployment measures intend to be independent from valid human intervention, to provide extra protection to individuals, especially in consideration of those systems falling outside of the scope of the GDPR's Article 22 provisions.[836] Alike to Article 22,[837] and the previously proposed testing requirements discussed above, those systems considered to be of a higher risk should be subject to a higher standard, to ensure stronger levels of protection for individuals, safeguard their rights, and encourage and promote the progression of ethical AI.

[835] The Information Commissioner's Office (n 718).
[836] GDPR (n 1) Article 22.
[837] ibid Article 22.

General Legislative Principles: Fundamental Concepts and Establishment of 'AI Human Oversight Board'

For a future framework for AI to work efficiently and succeed in encouraging and promoting the progression of ethical AI, it must introduce fundamental concepts that contribute to the foundation of the legislative measures. The following measures proposed include required levels of **transparency**, **explainability** and **interpretability** in law, in addition to how an establishment of a new authority could aid in matters of enforcement and redress. These proposed measures aim to strike a balance with strict regulation, that still compliments the progression of ethical AI.

Requirements for **transparency**, **explainability** and **interpretability** should be met at the initial development stage of the technology, and form central concepts of consideration for developers throughout. Similar to the recommendations previously suggested, it would be effective for a future framework to distinguish standards for systems that produce differing levels of effects, but a minimum level for all systems should still be set. The prescribed levels of these concepts need to work efficiently and successfully in practice, making it important to ensure developers can comply with the required standards. As previously discussed in respect to the suggestions offered by Bickerstaff and Mohan, [838] a public and accessible database should be established, to include the best approaches to satisfy the **transparency**, **explainability** and **interpretability** standards, and the most effective approaches to develop systems, to provide support to developers.

To ensure the strongest protection of human rights and promotion of ethical standards, it is crucial that any requirements set out are binding, and that all developers of AI systems comply. To emphasise the importance of developing AI to the ethical standards recommended, and to further influence compliance with such standards, a new authority should be established, to prevent adding to the current pressures and workload the ICO are already facing.[839] Across the literature, there is a general consensus which advocates for human oversight, but it is important to note that there do exist some limitations, and that effort should be made to ensure human

[838] Bickerstaff and Mohan (n 784).

[839] European Union Agency for Fundamental Rights (n 716) 10.

involvement is not a ‘rubber-stamping’ procedure, or a justification to argue that fundamental rights are being protected.[840]

A new authority directly focused on AI could provide security to the public by ensuring AI is developed ethically, and consequently, could improve public trust. A similar suggestion has been made within the AI Act, in which the EUAIB should help facilitate the implementation of the regulation with MS.[841] This thesis supports the establishment of an oversight board or similar in future regulation at a national level in England and Wales, to ensure that rules are complied with, whilst also encouraging best practices in the context of the culture and ethics of the country. The idea of a new independent body for AI has already reached Parliament, where the Artificial Intelligence (Regulation) Bill[842] proposes the creation of an AI authority, having a variety of functions to help address AI regulation. This idea is also one seen commonly in other fields, for example, the Copyright Hub, ICO, and Ofcom.

A future ‘AI Human Oversight Board’ could be granted responsibility to oversee the monitoring uploads from developers, and investigate any errors or flaws in algorithms to ensure they are rectified in a timely manner. Through investigation, the Board could provide recommendations on how to correct the flaw, and pose time restraints for when this needs to be achieved. If evidence of multiple or frequent flaws appear in systems, the authority could be granted powers to order independent reviews, in addition to the ability to carry out regular independent audits on systems, to ensure they are working as expected.

Further Considerations for Implementation

The suggestions for a future framework for AI intend to address the ethical and human rights challenges posed by AI systems. For ethical AI to be achieved, it must be ensured that systems are being developed in the correct ways, to not only prevent issues in the future, but also to comply with the proposed requirements that protect fundamental rights. Pre-deployment measures, in combination with post-

[840] Rikka Koulu, ‘Proceduralizing Control and Discretion: Human Oversight in Artificial Intelligence Policy [2020] 27(6) *Maastricht Journal of European and Comparative Law* 720.
[841] Provisional Agreement for the AI Act (n 103), Article 56.
[842] Artificial Intelligence (Regulation) Bill (n 101).

deployment measures, provide security and assurance to the public that systems are held to a high standard through all stages of their lifecycle, and that systems are being continuously monitored, rectified, and improved, therefore, consequently increasing the chances of increasing public trust in AI.

For such measures to work sufficiently in practice, AI systems need to be developed in a way that allows compliance with these requirements. The central concepts suggested in reference to the proposed new framework on AI intend to not only aid in developer compliance, but also for redress mechanisms to be readily available. In addition to this, a new authority could aid in the enforcement of the suggested requirements and provide guidance and support to developers. A new authority also could aid in making redress available and accessible. This chapter continues with further consideration of how these recommendations could aid the current problematic nature of placing liability in a scenario which includes an AI system and addresses the second limb of RQ1.

3.3 Accountability for AI Technology: Placing Liability

As noted within the literature,[843] AI brings complexities to the standard liability frameworks that currently exist, particularly due to the issue of foreseeability and the potential of unpredictability within systems. For ethical AI to exist and respect the fundamental rights and freedoms given to humans, the responsibility measures and processes must be made clear, considering the practicalities and possibilities of AI. For responsibility to be adequately and fairly placed, the concepts of **transparency** and **explainability** are imperative for the protection of individuals and their rights. As noted by Turner, although the current liability frameworks in the short term can be adjusted to address AI, the more important question is whether societal aims would be better served by "*reformulating the relationship with AI in a more radical fashion*".[844]

[843] European Commission, *Civil Liability – adapting liability rules to the digital age and artificial intelligence* (Inception Impact Assessment, 2021-22); European Parliament, *Artificial Intelligence and Civil Liability* (Policy Department for Citizens' Rights and Constitutional Affairs, 2020); Rodrigues (n 242); Cath (n 189); Hannah Sullivan and Scott Schweikart, 'Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?' [2019] 21(2) *AMA Journal of Ethics* E160.
[844] Turner (n 205) 132.

In light of this, this section of the chapter intends to answer the second limb of RQ1, and contribute by emphasising how a new framework for AI could add clarity, and more adequately address the liability issues that have arisen due to the rapid and widespread progression of AI. It is inevitable, that even with the strongest of safeguards and rules in place, that scenarios will still arise where AI systems act in unintended or unpredictable ways, due to the nature of the technology. For this reason, ethical principles and pre-emptive measures alone are not enough to sufficiently protect citizens and their rights,[845] nor to adequately promote and encourage the highest standard of ethical AI. Clear measures need to be established for AI systems post decision-making, to ensure that individuals can sufficiently challenge decisions, and that redress procedures are accessible, efficient, and provide adequate remedies.

For such redress measures to work in practice, further guidance is needed to clarify the relationship between AI systems and current liability frameworks. As seen in the literature, a series of questions arise in consideration of AI and current notions of liability, including whether AI is a product or a service,[846] whether a duty of care exists or should exist,[847] and whether AI technology should be judged on a standard similar to, or the same as the competency and expertise of humans.[848] In clarification of these questions, and ideally in a legislative and binding nature, redress procedures could be established to enable efficient routes to compensating individuals when necessary.

For individuals to challenge any decision made by an AI system, an understanding and explanation is needed as to how the decision was made, not only for challenges to be justifiable, but also for liability to be correctly placed according to related rules and guidance. The recommendations discussed in reference to **transparency** and **explainability** intend to work collaboratively, and provide a basis for other suggestions to ensure sufficient options and procedures of redress, to consequently

[845] Jess Whittlestone, Rune Nyrup, Anna Alexandrova and Stephen Cave, 'The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions' (University of Cambridge, 2019) <http://lcfi.ac.uk/media/uploads/files/AIES-19_paper_188_Whittlestone_Nyrup_Alexandrova_Cave.pdf> accessed 13th March 2021; Brent Mittelstadt, 'Principles alone cannot guarantee ethical AI' [2019] 1(11) *Nature Machine Intelligence*.

[846] Turner (n 205) 95.

[847] European Commission (n 67) 44.

[848] Abbott (n 402).

not only address the liability issues presented by AI, but also to provide further safeguards to society.

3.3.1 AI Technology and Possible Duties of Care

The Concept of a Duty of Care and Foreseeability

For a negligence claim to be successful, a plaintiff must prove that a duty of care exists between the parties in question, and that standard of care expected has been breached. The plaintiff must also prove that the lack of care has a causal link with the harm suffered and that the damage suffered was not too remote. A duty of care can be established through a 'special relationship' scenario, or through application of the 'neighbour principle'[849] established in common law. The principles of causation and remoteness reflect the link between the negligent conduct or omission and the harm suffered, and need to be established to ensure that there is a strong enough connection to fairly assign liability to the party that was negligent.

In the common law of negligence, several cases have explored the concept and scope of a duty of care. In one of the most infamous cases, the House of Lords in *Donoghue v Stevenson* established that manufacturers of a product owe a duty to consumers to take reasonable care not to cause harm, and introduced the 'neighbour principle' in reference to proximity.[850] This sentiment is one that is solidified in the more recent case of *Robinson*, whereby the Courts made clear that the well established principles of negligence should be applied to cases where possible, and extensions of these rules should only be considered in novel situations.[851] The standards of duties of care established throughout these cases are centred on the concept of foreseeability, which, as noted, poses a substantial issue in consideration of AI technology.

Due to the inherent notion of ML, evidence of AI acting in an unpredictable manner is already widespread.[852] Unpredictability in AI is defined as the inability to precisely

[849] *Donoghue v Stevenson* (n 59) [57] and [580].
[850] ibid [57] and [580].
[851] *Robinson v Chief Constable of West Yorkshire Police* [2018] UKSC 4 [2018] WLR 595 [30].
[852] Roman Yampolskiy, 'Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent' [2020] 7(1) *Journal of Artificial Intelligence and Consciousness* 109; Anne-Sophie Mayer, Franz Strich and Marina Fiedler, 'Unintended Consequences of Introducing AI Systems for Decision-Making' [2020] 19(4) *MIS Quarterly Executive* 239.

and consistently predict the specific actions a system will make to achieve its objectives, even in scenarios where the terminal goals of the system are known.[853] Examples include Google who claim their AI agent deliberately hid data from them to cheat at an appointed task,[854] and IBM's once hailed revolutionary cancer treatment, which then reportedly gave multiple examples of unsafe and incorrect treatment recommendations.[855] The unpredictability in AI technology stems from the ML capabilities within systems, in which behaviours are manifested only through interaction with the world and other agents in their environment.[856]

In respect of this, Hosanagar acknowledges that as AI becomes more intelligent and dynamic, the likelihood of unpredictability will increase.[857] Due to the likelihood and possibility of AI acting unpredictably, developers and organisations would be able to rely on the foreseeability concept that is needed for a successful negligence claim, and following the three-stage test,[858] would not owe a duty of care, so consequently, not be held liable. The principle of foreseeability has been subject to much consideration within the EU, in which it is made clear that foreseeability is required to ensure measures of protection against arbitrary interferences from public authorities, and so that individuals have the ability to adapt their behaviour and conduct to avoid penalties.[859] Hence, if individuals were held to be negligent for a decision which is deemed unforeseeable, this requirement set out by the ECtHR[860] would be undermined.

It should be noted that for 'special relationship' scenarios, the foreseeability concept is not as important. Certain duties, including a healthcare professional's duty to a

[853] Yampolskiy (n 806).
[854] Casey Chu, Audrey Zhmoginov and Mark Sandler, 'CycleGAN, a Master of Steganography' (NIPS Workshop 'Machine Deception', 2017) <https://storage.googleapis.com/gweb-research2023-media/pubtools/pdf/d8511d26602659849b93d28875f25780e37a973d.pdf> accessed 12th June 2020.
[855] Casey Ross, 'IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show' (Stat Blog, 25th July 2018) <https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/> accessed 18th January 2021.
[856] Roman Yampolskiy (n 806).
[857] Kartik Hosanagar, 'As machines become more intelligent, they also become unpredictable' (FoundingFuel Article, 2nd August 2019) <https://www.foundingfuel.com/article/as-machines-become-more-intelligent-they-also-become-unpredictable/> accessed 20th January 2021.
[858] *Caparo Industries plc v Dickman* (n 59).
[859] European Court of Human Rights, *Article 7: The quality of law requirements and principle of non-retrospectiveness of the criminal law under Article 7 of the Convention* (2019, Council of Europe Research Division) 9.
[860] *C.R. v the UK.* App No 20190/92 (ECtHR, 22nd November 1995) A/3550-C [33].

patient,[861] and a solicitor's duty to their client,[862] the duty of care is automatically established, based solely on the nature of the relationship. The standard of care in medical negligence, for example, has been developed through common law, and includes that a doctor must act in accordance with a responsible body of opinion,[863] and that it withstands logical analysis.[864]

Common law has also led to further clarification regarding the disclosure of risks in medical practice, and that such risks should be understood from a patient-centric view.[865] The concept of foreseeability, however, is still seen as a relevant concept to determine the placing of a duty being deemed fair.[866] It could be argued that a similar duty of care to disclose such risks should also be adopted, particularly due to the expertise of those deploying the technology. If such a duty were to be introduced for AI, it is essential that a clear and concise standard of the duty is given, and that it can work effectively in practice. The scope of such duties of care would have to be set out clearly within future legislation to reduce the pressure and give guidance to the courts.

The EC's report on the liability of AI suggests that legal personalities do not need to be granted to autonomous systems, and that any harm caused should be attributable to existing persons or bodies, advocating for a strict liability framework to address AI and the related liability issues.[867] Strict liability regimes, differing from notions of negligence which are usually hinged on fault or intention, instead allocate liability depending on the aspects of the act in question. To sufficiently protect individuals from the risk of unpredictability, errors and flaws within AI decision-making, binding guidance and rules need to be introduced to ensure efficient routes for all possible acts that may occur.

In relation to the unpredictability of systems, the report expresses that the requirements of equivalence be respected, and that use of a digital technology tool

[861] *Pippin v Sheppard* (1822) 147 ER 512.
[862] *Groom v Croker* [1939] 1 KB 194.
[863] *Bolam v Friern Hospital Management* Committee [1957] 1 WLR 582.
[864] *Bolitho v City and Hackney Health Authority* [1996] 4 All ER 771.
[865] *Montgomery v Lanarkshire Health Board* [2015] UKSC 11 [2015] 3 WLUK 306; *Chester v Afshar* [2004] UKHL 41 [2004] 10 WLUK 38.
[866] *Roe v Minister of Health [1954]* 2 WLR 915.
[867] European Commission (n 67) 37.

that has a degree of autonomy should not allow the operator to avoid liability, and instead, should give rise to liability to the same extent.[868] The principle of functional equivalence in this context refers to a scenario involving AI where compensation is denied, in which compensation would be offered in a functionally equivalent situation concerning conventional technology.[869] Due to AI being much more innovative in comparison to other types of technology, there is a clear need for binding measures and rules to address this current issue, and offer clear guidance to account for modern scenarios in which AI technology is used.

Explanations of AI Systems

As discussed, the concept of **explainability** in AI refers to systems having the ability to provide explanations for their decisions and is important for necessary oversight and testing to be achieved, to respect an individual's fundamental rights and freedoms.[870] Explanations of AI decisions can aid in detecting, and consequently correcting bias and inaccuracies, highlight errors or anomalies, and guarantee that machines are working as intended.[871] If demands for explainability were introduced within a future framework, consideration needs to be made to what constitutes an explanation, and the requirements for a sufficient explanation. Explainability demands should not be made in reference to the legal benefits alone, it also needs to be ensured that requirements made are technologically feasible in machines and must be written to enable a clear understanding to developers and those in the AI industry. It should also be acknowledged that for an explanation of a decision to be given, an understanding of the workings of models and the decision-making processes is also necessary.

[868] ibid 46.
[869] ibid 35.
[870] Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilovic, Karthikeyan Natesan Ramamurthy and Kush R. Varshney, 'TED: Teaching AI to Explain its Decisions' (IBM Research, 2019) <https://arxiv.org/pdf/1811.04896.pdf> accessed 12th January 2023; Alex Campolo, Madelyn Sanfilippo Meredith Whittaker, and Kate Crawford, *AI Now 2017 Report* (2017); Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, Jim Waldo, David Weinberger, and Alexandra Wood, 'Accountability of AI Under the Law: The Role of Explanation' (Berkman Center Research Publication, 2017, revised 2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3064761> accessed 12th January 2023; Goodman and Flaxman (n 195).
[871] Arrieta and Herrera (n 797).

Google's 'Vertex Explainable AI' offers tools and frameworks to aid in understanding and interpreting ML models, allowing deployers to debug and improve algorithmic performance, and help others understand the decisions produced.[872] It also offers solutions to explainability, in the form of encouraging systems that are designed with feature based and example based explanations embedded into them, in the hope to provide better understanding of model decision making, offering developers of AI a way to implement explainability into systems through code.[873] Another service available allows users to receive a score which explains how much each factor has contributed to a system's output, which could aid in verifying the system is working as intended and checking for bias and unfairness.[874] Ideas are also offered for ways to improve the model, and as such, this information could help in offering compensation and clarification to consumers, particularly in scenarios where consumers seek redress for a decision.

A general explanation can be defined as communication from one person, to another, in which justification is provided for an action or decision made.[875] A specific definition for system-to-human explanations is seen to be challenging,[876] however suggestions have been made in reference to this throughout the literature, motivated by the concept of 'meaningful information' introduced by the GDPR in reference to the right to explanation under Articles 13-15.[877] Hind et al introduce the Teaching Explanations for Decisions (TED) framework, and state that explanations should contain at least three elements; justification, complexity match and domain match.[878] They suggest that justification for a decision should increase trust in that decision made, and may include some information that can be verified by the consumer.[879] The complexity of explanations refers to the need for explanations to be produced in

[872] Google Cloud, 'Explainable AI' (Corporation Website) < https://cloud.google.com/explainable-ai > accessed 12th January 2022.
[873] ibid.
[874] ibid.
[875] Hind, Wei, Campbell, Codella, Dhurandhar, Moisilovic, Ramamurthy and Varshney (n 879).
[876] Finale Doshi-Velez and Been Kim, 'Towards a Rigorous Science of Interpretable Machine Learning' (2017) <https://arxiv.org/abs/1702.08608> accessed 15th January 2023.
[877] GDPR (n 1) Articles 13-15.
[878] Hind, Wei, Campbell, Codella, Dhurandhar, Moisilovic, Ramamurthy and Varshney (n 879).
[879] ibid.

an understandable form to the user's capabilities,[880] whereas the domain match indicates the incorporation of relevant terms in reference to the particular domain.[881]

Due to the lack of clear legislative standards to define a valid explanation, and consequently demand a valid explanation, the progress of AI continues to develop rapidly with little consideration for such components to be integrated into the machines themselves.[882] If a future framework were to introduce requirements to ensure sufficient explainability mechanisms were implemented into machines at the development and training stage, the procedures to challenge decisions would be put at ease.[883] It also needs to be considered whether one form or type of explanation would be deemed appropriate in all scenarios, and whether those systems that produce more substantial effects should be held to a stricter, or more detailed level of explanation in comparison to those of lower risks. Considering this, it can also be questioned whether a justified explanation is necessary in all scenarios, and if those systems that produce minimal consequences should be held to the same standard.

On a similar level, at times, particular AI systems can become too complex to be understood, or to provide explanations for decisions made.[884] For systems that are inherently unexplainable, often referred to as 'black box',[885] it needs to be considered what standard of legislative measures should be required. It should also be considered sensible to provide limitations to these systems, for example, not allowing unexplainable machines to produce decisions that have the potential to cause substantial effects. Although this would place a limitation that could stifle innovation to a degree, such a limitation would be arguably deemed proportionate in a human rights context.

[880] Amit Dhurandhar, Vijay Iyengar, Ronny Luss, and Karthikeyan Shanmugam, 'A Formal Framework to Characterize Interpretability of Procedures' (ICML Workshop on Human Interpretability, Sydney, 2017) <https://arxiv.org/abs/1707.03886> accessed 19th March 2022; Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong, 'Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' [2013] *Proceedings of IEEE Symposium on Visual Language Human-Centric Computing* 3.

[881] Hind, Wei, Campbell, Codella, Dhurandhar, Moisilovic, Ramamurthy and Varshney (n 879).

[882] ibid.

[883] Coeckelbergh (n 560).

[884] Information Commissioner's Office (n 518).

[885] Yampolskiy (n 806).

3.3.2 A Proposed New Framework for AI: The Recommended Liability Safeguards

For ethical AI to be achieved, it is imperative that a future framework clarifies the rules for liability in relation to the use and deployment of AI systems, and provides sufficient redress mechanisms. Liability principles must be 'fair' to all parties, and strike a balance between the importance of safeguarding individuals with encouraging and supporting developers, without posing restrictive or disproportionate stifling requirements. This would uphold the principle of proportionality, which has been richly referred to in case-law inside and outside of the UK, and most notably in the more recent case of *R(Bridges)*, where the use of FRT was deemed as disproportionate due to the police force not completing the appropriate policy requirements in assessing the risk to fundamental rights.[886]

This idea of proportionality has also been considered in the context of regulation, as seen by the Digital Services Act and its aim to create a safer digital space with a focus on the protection of end-user's fundamental rights.[887] This same approach underpins the EU's AI Act,[888] which regulates based on risk, creating regulation that is not unnecessarily burdensome, particularly for low-risk systems. To ensure proportionality and accountability, the establishment of an Oversight Board could be a key component. This Board is proposed to be responsible for monitoring systems, ensuring obligations are proportionate and necessary, and making sure that fundamental rights are adequately safeguarded. Such an Oversight Board could function similarly to other supervisory authorities, such as the ICO for the DPA, tasked with not only enforcing compliance but also providing guidance and redress mechanisms, and therefore reinforcing the principle of proportionality.

Due to the inherent nature of ML, allocation of liability cannot depend fully on the concept of foreseeability, suggesting rules in relation to strict liability may offer the best, or a more suitable alternative in certain scenarios. The introduction of new strict liability regimes have been discussed and advocated for throughout the literature,

[886] *R(Bridges) v CC of South Wales Police* (n 49); Equality Act 2010, s149(1).

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

[888] Provisional Agreement for the AI Act (n 103).

particularly in reference to systems that pose high risk effects.[889] Zech comments that strict liability for AI may incentivise further development of existing technologies, and in turn, aid in increasing and improving public acceptance.[890] The recommendations suggested consider the unpredictability of systems, to ensure that the allocation of liability is placed logically, and is dependent on the level of compliance with proposed established legislative duties.

Liability Allocation

As discussed, the common law precedents of duties of care would be difficult to apply to AI systems, particularly due to the foreseeability concepts included in the range of tests.[891] Due to this, the recommendations intend to create a legislative basis for a range of new duties of care for developers and deployers of systems. In the EC's report on the Liability of AI, an adapted range of duties of care were suggested for operators to comply with, including choosing the right system for the right task and skills, monitoring the system, and maintaining the system.[892] The Online Harms White Paper[893] proposed a duty of care on companies to regularly review the extent to which they are tackling harm and adapt their internal processes to drive improvement,[894] which could also be considered for developers of AI products.

In addition to this, the EC also advocate for duties to be imposed on developers to design, describe and market products, and on deployers to adequately monitor the product once put in circulation.[895] The thesis' recommendations support these suggestions by the EC,[896] in addition to imposing a duty on developers to comply with proposed testing requirements, and a duty on organisations to ensure that no unspecified or unjustified alterations are made to algorithms or systems.

[889] European Parliament, *Resolution of 20th October 2020 with recommendations to the Commission on a civil liability regime for Artificial Intelligence (2020/2014(INL))* (2021).
[890] Herbert Zech, 'Liability for AI: Public Policy Considerations' [2021] 22 *ERA Forum* 147.
[891] *Donoghue v Stevenson* (n 59); *Caparo Industries plc v Dickman* (n 59).
[892] European Commission (n 67) 44.
[893] Department for Digital, Culture, Media & Sport, and the Home Office, *Consultation outcome, Online Harms White Paper* (updated December 2020) Part 7.
[894] ibid Part 7.4.
[895] European Commission (n 67) 44.
[896] ibid 44.

In addition to this, the thesis calls for the imposition of a duty on those who deploy AI technology to disclose information about the AI system in use and the decision-making process, to ensure that sufficient options for redress are made available and accessible to individuals. To comply with this suggested obligation, deployers must offer accessible information that reflects the AI technology in use, the workings of the decision-making processes, and signposting for individuals who wish to challenge a decision that has been made. Such an obligation would not only provide security to individuals, and consequently improve public trust and consumer relations to benefit organisations, but also would raise awareness of AI, its purposes, workings and uses.

Claims that seek to challenge the developer's compliance to the suggested established duties would be decided upon the records and evidence stored and provided by the suggested new authority, or namely the recommended 'AI Human Oversight Board'. For scenarios in which systems have produced a decision that is deemed unexpected or unpredictable, the burden of proof could fall on developers to enable a limitation on their liability. Any alterations made to systems, whether that be by developers or organisations, should be justified before implementation, and should also be comparable to the standards recommended on the suggested online database that could list approaches to achieve **transparency**, **explainability** and **interpretability** in systems, in addition to processes to achieve the sufficient proposed levels of **accuracy**, **fairness**, and the mitigation of **bias**.

License and Compensation

In addition to these recommendations, consideration is also crucial for those systems or scenarios that do not fall within the prescribed rules, to ensure sufficient access to compensation to victims. A proposed establishment of a compensation fund for these scenarios intends to work in collaboration with the suggestions above, to limit liability for developers if they can prove a decision was unpredictable, and that all duties and requirements were sufficiently complied with. In scenarios where no fault can be established, the compensation fund should be used to ensure efficient and rapid options of redress to protect individuals and their rights.

The main question regarding compensation funds is where the capital should come from. Under Article 82 of the GDPR, data subjects have the right to claim compensation from data controllers or processors if material, meaning financial, or non-material damage, such as psychological damage, is suffered as a result of a breach of the provisions.[897] If controllers and/or processors disagree with claims for compensation, data subjects have to make a claim to the court,[898] which could deter individuals who do not have the finances, knowledge or confidence to enter the court process, particularly against major companies with international reputations. This system could also be considered unfair on small and medium sized enterprises, as such pay-outs could deplete funds owned by small or start-up businesses, not having the same effect on the larger companies. Such systems have the ability of retaining the monopoly of the 'big tech' organisations and restricting market competition.

Instead, a future framework could advocate for government to delegate some of its AI budget to a compensation fund used for breaches. A fund such as this could be used to reduce the liability on developers in scenarios where it may be deemed fair to do so. This would encourage and provide security to developers, consequently supporting further innovation and progression of ethical AI, as well as offering support to small and medium sized enterprises. In addition to government support, a future framework for AI could introduce a licensing scheme for certain systems, in which the funds accumulated are used within the widespread compensation fund, to reduce the hostility and pressure of data subjects arguing against international companies with more resources and funding. To ensure organisations still respect and adhere to such requirements, the licensing for their products could increase as a consequence of evidenced breaches, with increases made in line with the organisation's profits from the system, resources available and seriousness of the breach.

In scenarios including where AI systems produce decisions that have the potential to cause high risk effects, or where AI systems exist are inherently unexplainable, a

[897] GDPR (n 1) Article 82.

[898] Information Commissioner's Office, 'Taking your case to court and claiming compensation' (ICO Website) <https://ico.org.uk/for-the-public/data-protection-and-journalism/taking-your-case-to-court-and-claiming-compensation/> accessed 20th January 2023.

future framework could introduce a mandatory license for systems, in which rates are prescribed dependent on the system used, the processes and scores achieved at the testing and development stages of AI pre-deployment, and the related level of risk or lack of explainability. Mandatory widespread insurance schemes for AI have been suggested throughout the literature.[899] Compensation funds are argued to be an effective solution to protect victims, whose claims cannot be satisfied according to the applicable liability rules.[900] Such a scheme would also aid in conjunction where compulsory liability is introduced, whereby compensation can be given for redress caused by uninsured or unidentifiable technology, with reference to Article 10 of the Motor Insurance Directive[901] to exemplify such a scheme.

However, as noted by the EC, a widespread scheme alone would not aid in solving the problem of allocating liability, and cannot completely replace the clear and fair standards of liability rules.[902] For this reason, the rules on the licensing of specific AI technologies are suggested to work in collaboration with the recommendations discussed above, and made only available to systems in which decisions made are capable of producing high risk consequences and those that are inherently unexplainable. Licenses for specific systems, rather than all systems, ensure that the most restrictive measures in place are proportionate, and systems that would most likely benefit from the proposed compensation fund itself.

The related applications for licenses to deploy the mentioned specific systems could be made available after the successful completion of the DPIA requirement provided by the GDPR[903] and the additional suggested FRIA requirement. The cost and rate of licenses could depend on a range of factors stemming from the suggestions made by US Judge Karnow in reference to insurance schemes, including consideration of the potential risks of the related system.[904] In addition to this, factors such as the size and capabilities of the organisations deploying the systems, the necessity and

[899] Turner (n 205) 113; European Commission (n 67) 63; Karnow (n 376).

[900] European Commission (n 67) 62.

[901] Council Directive (EC) relating to insurance against civil liability in respect of the use of motor vehicles, and the enforcement of the obligation to insure against such liability (2009/103/EC) OJ L 263, Article 10.

[902] European Commission (n 67) 30.

[903] GDPR (n 1) Article 35.

[904] Karnow (n 376).

objectives of the decision-making process, and the use of systems could also be considered.

The proposed licenses could, for example, last for a maximum of five years before renewal, or until any substantial changes are made to systems. The funds accumulated by the licensing fees, in addition to the contributions from the state, could make up the proposed compensation fund, to be used for redress in scenarios where liability is unable to be established, or an unpredictable or unexpected decision has been proven to occur. If an organisation or developer of a licensed AI system is found at fault, or in non-compliance with the proposed duties, penalties may be given and struck against their license, whereby if a certain number of penalties have been given, the license is removed, and the system can no longer be legally deployed.

3.4 Conclusion

The recommendations suggested for a future framework for AI in this Chapter aim to work collaboratively to address the challenges posed to ethics, human rights and liability allocation by AI and its novel complexities. To achieve a successful future of ethical AI, fundamental principles and clear rules must form the basis of the proposed new framework for AI, to ensure the strongest protections to individuals and their rights. The requirements and obligations proposed, including the measures to ensure a sufficient standard of **accuracy**, **fairness** and **non-discrimination**, **transparency**, **explainability** and **interpretability** within AI systems, in addition to pre- and post-deployment duties to ease the allocation of fault and liability.

The establishment of a new authority, in addition to the proposed databases, licenses, redress measures and compensation schemes intends to provide support and guidance to developers, in addition to ensuring compliance with the proposed requirements, and to increase public acceptance, awareness, and trust of AI systems. This authority could follow a similar model to other regulatory authorities, such as the ICO, and perhaps be established with a multi-stakeholder structure, including experts in the field of AI, ethics, human rights and data protection. To ensure fairness, appointments could be made through an independent and transparent process, involving public consultation, and input from academia, industry representatives and members from governmental bodies. The Board should operate in cooperation with existing bodies that already exist, to ensure consistency, and to offer a holistic approach to AI regulation.

In combination with the recommendations to be proposed in Chapters Four and Five, these suggestions aim to collaboratively contribute by providing an effective, and extensive approach and perspective, to address the several complexities brought by AI to society. In Chapter Four, the conflicts that exist between AI and the GDPR[905] are analysed, in which further recommendations for a future framework for AI are introduced, in addition to suggestions in relation to reform of the DPA to further encourage and promote the progression of ethical AI.

[905] GDPR (n 1).

Chapter 4: AI and the GDPR

RQ2: *To what extent does the GDPR fail to account for the unique challenges posed by AI systems, and how could reform of the DPA drive more ethical AI use?*

4.1 Introduction

The purpose of this chapter is to use a doctrinal approach to examine the impact made by the GDPR on AI technology, and to evaluate the strengths and limitations of the GDPR provisions, namely the data protection principles,[906] and provisions that address automated decision-making, namely Article 22,[907] and Articles 13-15.[908] Building on the themes discussed in Chapter Three, which addressed the human rights and liability challenges posed by AI, this chapter continues by focusing on how the current rules on data protection interact with AI systems that process personal data. In doing so, the chapter aims to examine whether the GDPR is sufficient to address the ethical challenges highlighted in previous chapters, with consideration to transparency, fairness and accountability in AI. In reflection of this, the chapter aims to critically assess whether reform of the Data Protection Act[909] could strengthen protections for data subjects, aid further in promoting and encouraging ethical AI, and better align with the broader regulatory framework proposed within this thesis. In light of these findings, recommendations are proposed to provide stronger safeguards to data subjects, and to further promote and encourage the progression of ethical AI.

The GDPR, which came into effect in 2018, aims to protect the personal data of individuals within the EU,[910] however, the impact has been much more significant, starting a worldwide trend towards strengthening data protection.[911] Article 8 of the EU Charter on Fundamental Rights provides the right to the protection of personal data,[912] and the GDPR aims to uphold the protection of this right. Although the UK is

[906] ibid Article 5.
[907] ibid Article 22.
[908] ibid Articles 13-15.
[909] Data Protection Act 2018.
[910] GDPR (n 1).
[911] Lei Geral de Proteção de Dados (Federal Law no. 12,709/2018) (Brazil's Data Protection Act); Amendment of the Act on Protection of Personal Information 2020 (Act No. 57 of 2003 as amended in 2015) (APPI) (Japan's Data Protection Act).
[912] Charter of Fundamental Rights of the European Union (n 50) Article 8.

no longer explicitly subject to the GDPR due to Brexit,[913] the GDPR provisions were incorporated into England and Wales law through the Data Protection Act,[914] and therefore the same GDPR provisions with very minimal technical amendments, are still in effect today.[915]

Whilst the GDPR plays a pivotal role in setting data protection standards, it has notable limitations when applied to AI systems. The rapid development of AI and its technical complexities have highlighted gaps within the GDPR that were not initially addressed when the regulation was developed. The GDPR's core focus is on the processing of personal data, rather than the broader ethical and operational challenges that are posed by AI systems. Due to this, although the scope of the GDPR is substantial, it is insufficient to cover the entirety of AI's regulatory needs.

Following this trend in the recognition that data protection alone is insufficient to regulate AI effectively, the UK Government began a consultation on the developments of proposals to reform data protection standards post-Brexit, intending to give greater clarity over data protection rights.[916] In the consultation, it is acknowledged that AI development is contingent on data, and that data protection law, on its own, is unable to cover all possible AI use cases effectively.[917] This consultation led to a new Bill being introduced into Parliament,[918] showing the Government's intention of departing from the GDPR and regaining its independence to set itself apart from the EU. However, although the Bill intends to offer further clarity on data protection rights, alike to the GDPR, it focuses more on the issues pertaining to personal data rather than the broader regulatory challenges posed by AI systems.

As noted in the literature review, the GDPR definition of 'personal data' was carried forward from the previously existing DPD,[919] and includes any information relating to an identifiable natural person, who can be identified directly or indirectly, in particular

[913] Information Commissioner's Office, *Data Protection at the end of the transition period* (2019) 5.
[914] Data Protection Act 2018.
[915] Information Commissioner's Office (n 920).
[916] Department for Digital, Culture, Media & Sport, *Data: A New Direction- Government Response to Consultation* (June 2022).
[917] ibid.
[918] Data Protection and Digital Information Bill (No.2) HC Bill (2022-23, 2023-24).
[919] Data Protection Directive (n 465).

reference to an identifier such as a name, identification number, location data, or an online identifier relating to several listed factors.[920] This is consistent with the definition used within the new Data Protection and Digital Information Bill.[921]

As previously stated, AI technology is commonly trained and developed using multiple large datasets that contain personal data, and therefore is subject to GDPR provisions. With AI, additional data is better than less,[922] yet what is more important regarding this data, is that it is representative; a concept included within the recently approved AI Act.[923] However, this Article can be criticised due to the lack of clarity on how to achieve this, which is discussed further in Chapter Five.

As noted by the ICO, AI often works through the collection and analysis of all data available, rather than using a sample or segment of data.[924] Large amounts of data are essential for AI to achieve its full potential, whilst also helping to safeguard against errors, bias and unfairness.[925] As extensive amounts of data are essential towards achieving ethical AI, current and historical data protection safeguards to minimise the use of data face unprecedented challenges. The principle of data minimisation,[926] included as one of the key data protection principles in the GDPR, states that data controllers must ensure the processing of personal data is adequate to fulfil the stated purpose, relevant to that purpose, and limited to what is necessary.[927]

The Norwegian Data Protection Authority has noted that it is not possible to predict what some algorithms will learn, and that the intended purpose of algorithms is also likely to change and adapt as machines develop further.[928] This also poses a substantial challenge to the data protection principle of purpose limitation within the

[920] GDPR (n 1) Article 4(1).
[921] Data Protection and Digital Information Bill (n 925) cl 1.
[922] Christopher Kuner, Fred Cate, Orla Lynskey, Christopher Millard, Nora Loidesin and Dan Scantesson, 'Expanding the Artificial Intelligence- Data Protection Debate' [2018] 8(4) *International Data Privacy Law* 289; Viktor Mayer-Schönberger and Thomas Range, 'A Big Choice for Big Tech: Share Data or Suffer the Consequences' 97(5) *Foreign Affairs* 48.
[923] Provisional Agreement for the AI Act (n 103) Article 10(3).
[924] Information Commissioner's Office, *Big Data, Artificial Intelligence, Machine Learning and Data Protection* (2017, Version 2.2) 11.
[925] Kuner, Cate, Lynskey, Millard, Loidesin and Scantesson (n 929).
[926] GDPR (n 1) Article 5(1)(c).
[927] ibid Article 5(1)(c).
[928] The Norwegian Data Protection Authority (n 935) 18.

GDPR,[929] which states that personal data must be collected for specified, explicit and legitimate purposes, and that the personal data is not further processed in a manner incompatible with those purposes.[930] The data protection principles have clearly been made without strong consideration of the complexities of AI, and consequently, will cause further significant issues the more AI progresses.

The EU's AI Act and proposed Liability Directive[931] provide a more targeted framework for AI technology, producing specific provisions for AI systems, including transparency obligations and liability mechanisms that go far beyond the GDPR's provisions. For instance, the GDPR only directly addresses automated decision-making and profiling in regard to AI technology, and only in very few provisions, namely in Articles 13-15 on the right to explanation, explicitly for "*meaningful information about the logic involved*" to be given,[932] and in Article 22, which provides data subjects with a right not to be subject to decisions based solely on automated processing, including profiling, which produce legal or similarly significant effects.[933] In contrast, the detailed provisions within the EU's approach fill those gaps which are not sufficiently covered under the GDPR's current framework. For example, the literature is vast in debating the extent of existence to a right to an explanation within Articles 13-15,[934] in which a more clarified and explicit right is included within the non-binding Recital 71.[935] The terminology used within Articles 13-15 denotes a more general explanation of the process used,[936] rather than a specific explanation of the relevant decision made on the data subject included within the Recital.[937] This is just one example where the GDPR provisions are limited when applied to AI, where explainability is seen as an essential component for ethical AI.

As discussed in Chapter Three, it is essential that redress options are available to individuals, not only to promote and encourage ethical AI, but also to establish strong safeguards to protect fundamental rights. Understandable explanations to data

[929] GDPR (n 1) Article 5(1)(b).
[930] ibid Article 5(1)(b).
[931] Provisional Agreement for the AI Act (n 103) ; Proposed AI Liability Directive (n 104).
[932] GDPR (n 1) Article 13(2f), Article 14(2g) and Article 15(1h).
[933] ibid Article 22.
[934] ibid Article 13(2f), Article 14(2g) and Article 15(1h).
[935] ibid Recital 71.
[936] ibid Article 13(2f), Article 14(2g) and Article 15(1h).
[937] ibid Recital 71.

subjects must be made available in situations where rights have been affected through automated decision-making, not only to provide justification and reasoning for the decision, but to also ensure the decision that was made was accurate, fair, and free from bias. In the data protection context, consent to the processing of data is key, particularly when attempting to determine a lawful basis of processing.[938] The literature has noted the issue on the current rules in gaining consent, and how this is often bypassed,[939] which leads to the concept of future consent, and how if an individual consents to the processing of their data on one occasion, this does not necessarily mean they consent in the future.

Article 22 of the GDPR has a direct focus on automated processing, including profiling, and aims to provide a safeguard to data subjects who are subject to decisions based solely on automated decision-making.[940] As noted in the literature review, this Article has a particularly limited and vague scope, with no explicit or clear definition of the term ‘based solely’, and a lack of clarification in relation to the meaning of ‘similarly significant effects’.[941] The Government’s consultation also picks up on these issues, acknowledging the concern towards the Article being “*futureproof*”.[942] Article 22 also provides three exceptions to the provided right, in the cases of contractual, and legislative measures, and where the data subject has explicitly consented.[943] These exceptions are notably broad, which gives rise to concerns with data controller and company misuse to bypass the limiting provision of Article 22. In light of the listed exceptions, Article 22 also provides that data controllers should implement ‘suitable measures’ to safeguard individual rights and freedoms, namely at least, the right to obtain human intervention, to express their point of view and to contest the decision.[944] These measures are also noticeably vague, and ideally need further clarification to provide stronger safeguards to individuals, more specific and clearer rules for data controllers, and to further promote and encourage the progression of ethical AI.

[938] ibid Article 6.
[939] Adam Andreotta, Nin Kirkham and Marco Rizzi, ‘AI, Big Data, and the Future of Consent’ [2021] 37 *AI & Society* 1715.
[940] GDPR (n 1) Article 22(1).
[941] ibid Article 22(1).
[942] Department for Digital, Culture, Media & Sport (n 923).
[943] GDPR (n 1) Article 22(2).
[944] ibid Article 22(3).

Continuing, this chapter seeks to critically examine the relationship between the GDPR and AI, to state which areas of the GDPR could benefit from further clarity. In light of these findings, recommendations are proposed in relation to how potential reform of the DPA[945] could benefit AI technology, and where a proposed new framework for AI could aid in connecting and filling the gaps within the GDPR, and provide more specific rules and guidance for AI systems.

4.2 The Relationship between AI and the GDPR

4.2.1 Personal Data, Sensitive Data, and Inferences

As previously stated, the GDPR covers the processing of all personal data,[946] and due to personal data usually being part of large datasets, the majority of AI automated processing is subject to the GDPR. ‘Personal data’ is not exhaustively defined, with the GDPR providing the arguably wide and vague scope of any information relating to an identified or identifiable natural person.[947] An identifiable natural person is described to be one that can be identified, directly or indirectly, through an identifier such as a name, identification number, location data, online identifier, such as an email address, or factors specific to the physical, psychological, genetic, mental, economic, cultural or social identity of that particular person.[948] This lack of an exhaustively defined list has also led to questions and disputes relating to the scope of the right to information, particularly concerning processing made by automated machines, and the potential inferences drawn from processing.[949]

The GDPR, like several other data protection laws, distinguishes between general personal data, and ‘special categories’ of personal data. Within the GDPR provision, Article 9 specifies these special categories as personal data that reveal racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership.[950] The processing of genetic data, health data, biometric data for the purpose of identifying a person, or data relating to a person’s sex life or sexual

[945] Data Protection Act 2018.
[946] GDPR (n 1).
[947] GDPR (n 1) Article 4(1).
[948] ibid Article 4(1).
[949] Stefanie Haenold, ‘Profiling and Automated Decision-Making: Legal Implications and Shortcomings’ in Marcelo Corrales, Mark Fenwick and Nikolaus Forgó (eds), *Robotics, AI and the Future of Law* (2018) 143.
[950] GDPR (n 1) Article 9(1).

orientation is also included within the scope of 'special' personal data.[951] Article 9 of the GDPR prohibits the processing of these special categories of personal data, but provides an extensive list of exceptions, including scenarios where the data subject has given explicit consent, for reasons of substantial public interest including public health, and where processing is carried out in the course of its legitimate activities with appropriate safeguards.[952]

As acknowledged by Ufert, although the GDPR gives control to data subjects on how their personal data is collected and processed, very little control is given on how the data is consequently evaluated, and used to draw inferences about the data subjects.[953] This has also been evident through the historical development of common law, in which the courts have held that in scenarios where a data subject wishes to challenge evaluations or analysis of their personal data, sectoral laws relating to the circumstances of the case are relied upon, rather than existing data protection laws.[954] However, the ICO comments that the degree of certainty of the inference, and whether the inference is being deliberately drawn, should be considered when deciding the scope of 'sensitive' personal data.[955]

This conflict was also reflected in case-law, whereby in *YS and Others*,[956] the analysis of personal data was not in itself classified as personal data, whereas in *Nowak*,[957] a broader scope of personal data was accepted, in the form of opinions and assessments, although only in certain circumstances and based on a case-by-case assessment. Clifford, Richardson and Witzleb highlight the new challenges posed by AI, or more specifically, ML developments, that have capability of producing 'sensitive' inferences out of data that may not originally considered to be 'sensitive',[958] and therefore not subject to the stronger safeguards. In light of this, Ufert notes that the limited rights of data subjects over potential inferences become

[951] ibid Article 9(1).
[952] ibid Article 9.
[953] Ufert (n 4).
[954] Case C-28/08 *Commission v Bavarian Lager* [2010] ECR 2010 I-06055 [49-50].
[955] Information Commissioner's Office, 'Special Category Data' (ICO Website and Guidance) <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-is-special-category-data/#scd7> accessed 2nd March 2021.
[956] Joined Cases C-141/12 and C-372/12 (n 467) [45-47].
[957] *Nowak v Data Protection Commissioner* (n 469) [54-55].
[958] Damian Clifford, Megan Richardson and Normann Witzleb, 'Artificial Intelligence and Sensitive Inferences: New Challenges for Data Protection Laws' in Mark Findlay, Jolyon Ford, Josephine Seoh and Dilian Thampapillai (eds) *Regulatory Insights on Artificial Intelligence: Research for Policy* (2021, Edward Elgard) 2.

substantially problematic when processed using AI technology in this way, particularly with systems capable of de-anonymising data sets in which such inferences from that data are used to make substantial decisions in relation to data subjects.[959]

This has led to a recent dispute in the CJEU on the scope of Article 9 of the GDPR on special category data[960] in relation to inferences. In *OT v Vyriausioji tarnybinės etikos komisija*,[961] the CJEU chose to interpret the scope of the Article broadly, after Lithuanian law ordered individuals who received public funds to make declarations of interest, including the name of their spouse, cohabitee, or partner.[962] These declarations were published online, and it was found that inferences could be taken from this published information to indirectly determine one's sexual orientation (a special category data under Article 9)[963]. Due to this, the rules for special category data applied, and the processing as such was prohibited under Article 9, subject to the usual exceptions.[964] This case has been commended for providing clarity on a point previously contrasted in EU MS case-law,[965] and reflects the strength of data protection regulation, even if personal data is identified indirectly.

The ICO highlights that although conditions to minimise privacy risks should be introduced where potential inferences on 'special' personal data could be made, especially where the inference is not relevant to the technologies purpose, the DPA does not include such a requirement.[966] Strengthening safeguards in relation to the processing of 'sensitive' personal data as reflected in *OT v Vyriausioji tarnybinės etikos komisija was clearly* warranted due to the rise of heightened legitimate fears of sensitive data misuse.[967] AI systems also could possibly make unreasonable or

[959] Ufert (n 4).
[960] GDPR (n 1) Article 9.
[961] Case C-184/20 *OT v Vyriausioji tarnybinės etikos komisija* [2022] ECLI:EU:C:2022:601.
[962] ibid.
[963] GDPR (n 1) Article 9.
[964] GDPR (n 1) Article 9(2).
[965] Datatilsynet, 'Administrative Fine – Grindr LLC' (Norwegian DPA imposes fine, 2021) <https://www.datatilsynet.no/contentassets/8ad827efefcb489ab1c7ba129609edb5/administrative-fine---grindr-llc.pdf > accessed 12th January 2023; Agencia Española Protección datos, 'Grindr LLC' (Spanish DPA Decision, E/03624/2021) <https://gdprhub.eu/index.php?title=AEPD_(Spain)_-_E/03624/2021> accessed 12th January 2023.
[966] Information Commissioner's Office (n 961).
[967] Ufert (n 4); Alexandre Veronese, Alessandra Silveira and Amanda Lemos, 'Artificial Intelligence, Digital Single Market and the proposal of a right to fair and reasonable inferences: a legal issue between ethics and techniques' [2019] 5(2) *EU Law Journal* 75.

incorrect inferences, which consequently could have a harmful impact to individuals, and without sufficient protections, could potentially lead to unfair and unappealable decisions.[968] For this reason, it is imperative that future legislative provisions make a full consideration of the complexities of AI, and that when necessary, provide explicit clarity on the novel technical capabilities that have not needed the same depth of consideration previously.

4.2.2 The GDPR Provisions in Need of Clarity

As noted in the literature review, the rapid progression of AI technology poses challenges to provisions within the GDPR, giving rise to whether these provisions need further clarification, or if new provisions need to be introduced. For ethical AI to be promoted effectively, the scope of legislative provisions must be able to deal with the novel complexities of AI sufficiently, and with clarity. Although the literature supports the view that currently, the GDPR does not fulfil this criterion, less consideration is made towards how explicit provisions can be improved or adjusted to reduce these concerns. In light of this, the UN has also highlighted the potential issues brought to the traditional data protection concepts of consent, transparency, purpose and use, and accountability, highlighting these as the pillars upon which international data protection is founded.[969]

Supporting this, Spyridaki comments that currently, the GDPR complicates the processing of data in several ways in an AI context,[970] and the UN has suggested that data protection regulation must be updated, to account for the new challenges posed to privacy brought by AI.[971] Kuner et al note that protecting data privacy is more important than ever given the impact, speed, and difficulty in analysing and explaining AI, and highlight the importance of expanding the focus of debate from mere compliance, to enhanced safeguards and effective governance.[972] It must also be ensured that the principles of accessibility and foreseeability are upheld sufficiently, and that if provisions are too broadly defined, this may prevent

[968] Rania El-Gazzar and Karen Stendal, 'Examining How GDPR Challenges Emerging Technologies' [2020] 10 *Journal of Information Policy* 237.
[969] United Nations General Assembly (n 82) [35]; United Nations Human Rights Committee (n 479) [10].
[970] Spyridaki (n 84).
[971] United Nations General Assembly (n 82) [35]; United Nations Human Rights Committee (n 479) [10].
[972] Kuner, Cate, Lynskey, Millard, Loidesin and Scantesson (n 929).

assessment of whether provisions are capable of lawfully infringing on rights in proportionate ways.[973] This also aligns with respect for the rule of law, which encourages legal rules to be clear and precise,[974] to ensure that actions made under legislation are prescribed by law.[975]

In respect of this, the next section of this chapter (4.3) identifies and explores three areas of the GDPR that are viewed to be insufficient in dealing with all aspects of AI effectively. To begin, a critical examination is made of the data protection principles contained within Article 5,[976] with full consideration of the technical abilities of AI, and the potential challenges posed by these principles. In addition to this, the debate regarding the 'right to explanation' within Articles 13-15[977] will be explored, to offer suggestions towards providing further clarity and guidance to AI developers. Article 22,[978] which provides data subjects the right not to be subject to a decision based solely on automated processing is also assessed, with recommendations introduced to reduce the vagueness and lack of clarity of the provision's scope.

Although there have been concerning reports from the Government to remove Article 22 completely, after a vast majority of respondents opposed this proposal, the Government is now considering how to amend and clarify the circumstances in which it will apply.[979] Although further guidance on the GDPR is imperative for providing clarity to developers of AI, eventually, it is inevitable that new provisions will need to be introduced, to deal with the specific complexities and to further encourage ethical AI to a more sufficient standard.

4.2.3 Using the GDPR as a Basis for a Proposed New Framework for AI

The rapid progression of AI in modern history has already taken technological limits to a new level, making a substantial impact, and in several cases, an improvement in the efficiency of several industries. Concern has always existed concerning the unregulated progression of AI, and this will continue to grow as AI reaches higher

[973] *Szabó and Vissy v. Hungary*. App No 37138/14 (ECtHR, 12 January 2016).
[974] Council of Europe, *Report on the Rule of Law* (Venice Commission, 2011) 11.
[975] Joined cases C-203/15 and C-698/15 (n 838) AG Opinion [137-154].
[976] GDPR (n 1) Article 5.
[977] ibid Articles 13-15.
[978] ibid Article 22.
[979] Department for Digital, Culture, Media & Sport (n 923).

levels of intelligence.[980] Within the UK, several soft law approaches, guidance, and proposals to address AI have been introduced in various industries,[981] however, there is no hard legislation focused directly on the specific aspects of AI and the challenges posed. As previously stated, the GDPR only addresses automated decision-making and profiling explicitly, but is interpreted to cover a wider range of AI systems under its provisions dependent on the processing nature.

As discussed in the next section of this chapter (4.3), the extent of the effectiveness of the GDPR provisions can be questioned, and is often argued to be too general and vague to sufficiently deal with the complexities of AI.[982] It is inevitable that new challenges and issues will come to light as AI continues to progress, making it imperative that the scope of future legislation is capable of dealing with such issues in a practical and direct way, whilst effectively promoting ethical AI. For this reason, this thesis advocates for a new legislative framework focused on AI, which uses the GDPR as a foundation to its structure. This is an approach touched upon by the UN in their report on the promotion and protection of the right to freedom of opinion and expression, which makes the estimation that robust data protection legislation that addresses AI-related concerns may be considered by states.[983]

The need for a new framework specifically for AI is evidently supported by the EU, who have released the White Paper on AI for 'high risk' systems,[984] which led to the recent approval of the AI Act.[985] As noted by Van Ooijen and Vrabec, current data protection laws, such as the GDPR, rely upon assumptions of human decision-making,[986] rather than having a specific focus on the complexities of AI. Wachter, Mittelstadt and Floridi acknowledge that several algorithmic decisions are outside of

[980] Cambridge Consultants (n 4) 63; AI Now (n 260) 8; Scherer (n 70); Borgesius (n 4); United Nations Human Rights Committee (n 479) [10].
[981] Department of Health and Social Care, 'New Code of Conduct for AI systems used by the NHS' (Research and Innovation, Government website, 19th February 2019) <https://www.gov.uk/government/news/new-code-of-conduct-for-artificial-intelligence-ai-systems-used-by-the-nhs> accessed 5th March 2021; Information Commissioner's Office (n 208); Information Commissioner's Office (n 98); European Commission (n 112); European Commission (n 67).
[982] Borgesius (n 4); Ufert (n 4); Spyridaki (n 84).
[983] United Nations (n 82) [45] and [63].
[984] European Commission (n 112).
[985] Provisional Agreement for the AI Act (n 103).
[986] I. Van Ooijen and Helena Vrabec, 'Does the GDPR Enhance Consumers' Control over Personal Data? An Analysis from a Behavioural Perspective' [2019] 42 *Journal of Consumer Policy* 91.

the GDPR's scope and therefore its rules,[987] reflecting limited application in relation to AI.

This gives rise to the idea that AI may benefit from a separate framework, not only to avoid complexities and conflicts within the GDPR, but also to focus on the inherent challenges brought by AI, and to consequently address them to a more sufficient standard. In support of this view, Winston comments that the GDPR can serve as a complementary basis to AI regulation, and that each regulation should not be exclusive of one another.[988] Spyridaki agrees, expressing that the impact of the GDPR on the data market cannot be ignored, and therefore regulation and the GDPR should be considered together, to ensure consistency.[989] The Centre for Information Policy Leadership also support this, highlighting that for ease of practical implementation by developers, and to avoid unnecessary duplication or potentially conflicting provisions, future AI regulation needs to take into consideration the GDPR.[990]

These suggestions not only reflect the acceptance that AI will need a separate specific framework to sufficiently address the unique challenges brought, but also the importance of the GDPR when considering future proposals. Using this, the recommendations proposed in this thesis use the GDPR as a foundation and inspiration towards new legislative regulation of AI. Using the GDPR as a basis to a new framework will ensure not only consistency between the two regulations, but also allow a more ethical and focused scope on AI, addressing the various challenges to the technology with a higher, more sufficient, and practical standard. For AI to develop and progress ethically and effectively, it is imperative that public trust is increased, both through safety and liability aspects, but also that individual's fundamental rights are sufficiently protected.

In a study conducted in 2018 by Mohallick et al in relation to 'recommender systems', which commonly integrate AI profiling, the majority of respondents argue that these

[987] Wachter, Mittelstadt and Floridi (n 195).
[988] Winston (n 85).
[989] Spyridaki (n 84).
[990] Centre for Information Policy Leadership (n 86) 19.

systems violate user privacy.[991] Although user trust may be difficult to re-establish, it is essential that future legislative rules allow trust to naturally increase and improve over time, by providing clarity to developers, in addition to strong and effective safeguards to protect rights, freedoms and interests. In addition to this, new legislation specifically for AI gives rise to the opportunity to express a clear aim and purpose for such legislation, namely, to effectively promote and encourage ethical AI. Discussed further in Chapter Six, a more specified and clear purpose expressed for focused legislation would also allow the explicit concerns in relation to AI technology to be addressed, for example, by requiring data controllers and data processors to make more consideration to the societal impact regarding the use of AI systems.[992]

4.3 The GDPR, and Future Reform Recommendations

4.3.1 The GDPR Data Protection Principles

Article 5 of the GDPR governs the principles relating to the processing of personal data, and includes 'lawfulness, fairness and transparency', 'purpose limitation', 'data minimisation', 'accuracy', 'storage limitation', 'integrity and confidentiality', and finally, 'accountability'.[993] The data controller must ensure that compliance with these principles is demonstrated. As previously noted, although these principles are essential for general data protection, to sufficiently address the complexities of AI in full, a future framework for AI legislation needs to provide a refocused scope in relation to these principles. As stated within Chapter Three, for the purposes of effectively encouraging and promoting ethical AI, such refocused principles can take inspiration from Article 5 of the GDPR,[994] but be developed further to deal with the specific capabilities of AI systems. Some of the data principles, namely 'purpose limitation' and 'data minimisation' may need to be adjusted for when automated decision-making is used.

[991] Itishree Mohallick, Katrien De Moor, Özlem Özgöbek and Jon Gulla, 'Towards New Privacy Regulations in Europe: Users' Privacy Perception in Recommender Systems' in G. Wang, J. Chen and L. Yang (eds) *Security, Privacy and Anonymity in Computation, Communication, and Storage* (2018, Springer).

[992] Nóra Loideain and Rachel Adams, 'From Alexa to Siri and the GDPR: The gendering of Virtual Personal Assistants and the role of Data Protection Impact Assessments' [2020] 36 *Computer Law & Security Review* 105366; Lorena Jaume-Palasi, 'Why Are we Failing to Understand the Societal Impact of Artificial Intelligence' [2019] 86(2) *Social Research: An International Quarterly* 477.

[993] GDPR (n 1) Article 5.

[994] ibid Article 5.

The data protection principle relating to 'purpose limitation' within the GDPR states that personal data must be collected for specified, explicit and legitimate purposes, and not further processed in a manner incompatible with the stated purposes.[995] As previously noted, the specified purpose of AI is likely to change and adapt as it develops further,[996] posing a clear challenge to the purpose limitation principle. However, this principle also includes some exceptions, namely in scenarios where further processing is for archiving purposes in the public interest, scientific or historical research purposes, or for statistical purposes.[997] There is no further explanation or definition as to what is considered under the scope of this exception, which could include AI development within 'scientific purposes'.

The data protection principle that relates to data minimisation, as previously noted, poses a challenge to AI technology, which ideally, needs to be trained using the most data available, to improve **accuracy**, **fairness,** and **non-discrimination** within machines. If ethical requirements and obligations set out in a new legislative framework were to be met by data controllers, the restrictive nature of the 'purpose limitation' and 'data minimisation' principles could eventually be removed, to allow AI to develop to the best of its abilities, and aid in achieving the promotion of ethical AI. These matters are explored and evaluated further in Chapter Six, which sets out the recommendations for a proposed new framework for AI.

4.3.2 Articles 13-15 of the GDPR, and the Debated 'Right to Explanation'

Articles 13 and 14 of the GDPR govern the data subject's right to be informed, namely for information to be provided where personal data is collected from the data subject,[998] and for information to be provided where personal data has not been obtained from the data subject respectively.[999] Article 15 provides the right of access by the data subject, namely that the data subject shall have the right to obtain confirmation from the data controller as to whether or not personal data concerning

[995] ibid Article 5(1)(b).
[996] The Norwegian Data Protection Authority (n 543) 18.
[997] GDPR (n 1) Article 5(1)(b).
[998] ibid Article 13.
[999] ibid Article 14.

them is being processed.[1000] As previously noted in Chapter Three, an obligation on data controllers is established within Articles 13-15 that specifically addresses automated decision-making. This obligation requires data controllers to provide data subjects with meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing, in scenarios where there is automated decision-making.[1001] Throughout academia, these provisions have led to the extensive debate on whether a 'right to explanation' exists within the GDPR.

As previously noted in Chapter Three, **explainability** is essential for encouraging and promoting ethical AI. Wachter et al argue that a provided right to explanation does not exist within Articles 13-15 of the GDPR, and that the lack of explicit, well-defined and precise language runs the risk of the safeguards being ineffective.[1002] A stronger and more valid example can be seen with the paired Recital on profiling.[1003] Recital 71 of the GDPR states that processing should be subject to suitable safeguards, including obtaining an explanation of the decision reached after such assessment, and an opportunity to challenge that decision.[1004] As stated by the ECJ, whilst Recitals may aid in the interpretation of given legal rules, they cannot in itself constitute such a rule.[1005]

The terminology used within Recital 71[1006] provides a much stronger and clearer right in comparison to the terminology used in Articles 13-15,[1007] which suggests a more general explanation of the systems used. However, the Article 29 Working Party has stated that although the controller does not necessarily need to give a "*complex explanation of the algorithms used or disclosure of the full algorithm*"[1008] to conform to the rules within Article 13-15, the scope of 'meaningful information'

[1000] ibid Article 15.
[1001] ibid Article 13(2f), Article 14(2g) and Article 15(1h).
[1002] Wachter, Mittelstadt, and Floridi (n 195).
[1003] GDPR (n 1) Recital 71.
[1004] ibid Recital 71.
[1005] Case 215/88 *Casa Fleischhandels-GmbH v Bundesanstalt für landwirtschaftliche Marktordnung* [1989] (ECR-2789) [31].
[1006] GDPR (n 1) Recital 71.
[1007] ibid Articles 13-15.
[1008] Article 29 Working Party (n 494) 25.

should be "*sufficiently comprehensive*",[1009] to ensure that data subjects gain an understanding of the decision.[1010]

The precise scope of the right to explanation has also been examined within the courts, where in the cases of *Uber Drivers v Uber (transparency requests)*[1011] and *Ola Drivers v Ola*,[1012] the courts confirmed that if automated decision-making within the scope of Article 22 has been used, the claimants would be able to access 'meaningful information about the logic involved' in these algorithms.[1013] In the Uber case however, the court decided that Article 22 was not invoked, and hence, Uber were not obliged to give this information.[1014] In the Ola case, the court held that only the automated system of 'penalties and deductions' fulfilled the scope of Article 22,[1015] as such penalties produced 'similarly significant effects' to the claimant, and hence, Ola had to communicate the 'main assessment criteria', in addition to 'their role in the automated decision' to allow the claimants to understand the basis of the decisions made.[1016] These cases reveal the importance of clarity when assessing the application of Article 22, as when applied, the 'right to explanation' under the GDPR is triggered.

The significance of Article 22 in these cases is not to be underestimated. However, the Future of Privacy Forum has argued that meaningful information should be provided, even in circumstances where a situation is not within the scope of Article 22.[1017] This has been the interpretation within Austria, where the data protection authorities had held that data controllers must adhere to the provision of meaningful information, even if the decision-making falls outside of the scope of Article 22.[1018] Following this decision, the 'meaningful information' disclosed amounted to the input variables, parameters, the effect these had on the decision, and why the data subject

[1009] ibid 25.
[1010] ibid 25.
[1011] Uber v Uber Drivers (Transparency Case) [2021] C /13/687315/HA RK 20-207
[1012] *Ola v Ola Drivers* (n 529).
[1013] *Ola v Ola Drivers* (n 529) [3.1]; *Uber v Uber Drivers (Transparency Case)* (n 1017) [3.1].
[1014] *Uber v Uber Drivers (Transparency Case)* (n 1017) [4.66-4.67].
[1015] *Ola v Ola Drivers* (n 529) [4.51].
[1016] ibid [4.52].
[1017] Future of Privacy Forum, *Automated Decision-Making Under the GDPR* (May 2022) 25.
[1018] Datenschutz Behörde, 'Case DSB-2020.0.436.002' (Austria DPA Decision, 2020) <https://gdprhub.eu/index.php?title=DSB_(Austria)_-_2020-0.436.002> accessed 20th January 2023.

had received the particular score, rather than information that would be considered trade secrets, like disclosure of the algorithm itself.[1019]

Considering the above discussion, this thesis advocates for a strengthened right to explanation to data subjects, similar to what is suggested within Recital 71. This would make it clear that the explanation given must correlate to an individual past decision, to ensure clarity and preciseness, like the interpretation given in Austria.[1020] Recommendations are proposed in light of this below, in addition to the extent to which explainability is possible in AI, and how this can be achieved to ensure developers can satisfy the suggested adjustments to legislative requirements. It is also important to note that these recommendations will only become more effective once Article 22 has been clarified, which is the next point of discussion that follows within this chapter.

A Proposed Adjustment of Articles 13-15 of the GDPR: To Increase Public Trust concerning the Obligation Relating to the Provision of 'Meaningful Information about the Logic Involved'.

To effectively encourage and promote ethical AI, this thesis advocates for a change to the terminology contained within Articles 13-15 that explicitly refers to automated decision-making. As previously noted, Articles 13-15 of the GDPR provide an obligation for data subjects to be given meaningful information about the logic involved in decisions, at the time such personal data is obtained by data controllers, to ensure fair and transparent processing.[1021] In line with the suggestions proposed in Chapter Three, general information about the automated decision-making processes should already be made readily available and accessible, to promote what is being done by organisations to ensure sufficient protection of fundamental rights.

To eventually achieve ethical AI, public trust and awareness of AI must be improved. It is imperative, to increase awareness, educate society, and eventually increase public trust in AI technology, that data controllers are transparent regarding the automated processing they use, how it works, and the impact on individuals. For this

[1019] ibid.
[1020] ibid.
[1021] GDPR (n 1) Article 13(2f), Article 14(2g) and Article 15(1h).

reason, as stated in Chapter Three, future legislation must place an obligation on data controllers, to ensure such information is made available to the general public and is easily accessible, but ideally also openly advertised. This adjustment would also effectively support the promotion of ethical AI, as if such an obligation existed, a level of competition is likely to arise between data controllers regarding the quality of the information provided. In addition to this, competition is likely to form between manufacturers of AI, who would be actively encouraged to ensure such information could be made available, consequently improving the ethical development of AI in the process.

A Broader Legislative Requirement of a Right to Explanation: Stronger and Clearer Safeguards to Protect Individual's Rights and Freedoms.

In terms of the explainability of machines, and in reference to the terminology used within both Articles 13-15 and Recital 71 of the GDPR, two different types of explanation can be given. These include an explanation of the system's functionality and the processes used, as described above, and explanations about a past decision, as suggested by the Article 29 Working Party[1022] and in the *Uber* and *Ola* cases.[1023] Wachter et al distinguish between these two types of explanation, making it clear that system functionality refers to the logic, significance, envisaged consequences, and general functionality of a system, in comparison to past decisions, in which the rationale, reasons and context of a specific decision are included.[1024] For rights and freedoms to be sufficiently protected and upheld, it is essential that a requirement exists in legislation that demands explanations be made available of a past decision that has been challenged, so that it can be sufficiently contested when necessary.

For redress options to work effectively, individuals need to be able to understand the rationale behind a decision being made on them, to have the ability to contest it. For this reason, this thesis advocates for a more developed right to explanation to be considered regarding past decisions made by AI. This suggestion is more aligned

[1022] Article 29 Working Party (n 494) 25.
[1023] *Ola v Ola Drivers* (n 529) [3.1]; *Uber v Uber Drivers (Transparency Case)* (n 1017) [3.1].
[1024] Wachter, Mittelstadt, and Floridi (n 195).

with the current right to explanation provided in the non-binding Recital 71,[1025] in addition to what was stated within the drafting of the GDPR. For such a right to be introduced, the explanations demanded must be technically and legally feasible, and guidance must be produced not only to ensure compliance, but to aid in promoting the ethical development of AI. To ensure this, the proposed legislative framework for AI aims not only to set high standards to provide strong safeguards to individuals, but to also ensure data controllers can easily comply with the obligations suggested.

4.3.3 Article 22 of the GDPR

Article 22 of the GDPR is one of the few provisions that specifically addresses automated decision-making and profiling.[1026] The Article provides the data subject the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal or similarly significant effects.[1027] 'Profiling' is defined within the GDPR as any form of automated processing of personal data to evaluate certain personal aspects relating to a natural person, in particular, to analyse or predict aspects relating to work performance, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.[1028]

It should be noted that the recently introduced Data Protection and Digital Information Bill seeks to depart from Article 22, reframing the provision to generally allow automated decision-making, with the Article 22 'safeguards' applying only where special category data is concerned.[1029] Other than this difference, the same phrasing of Article 22 is used, so this assessment of Article 22 can also be applied to the wording within the Bill. It should also be highlighted the precarious approach taken by Parliament in attempting to reduce the Article 22 safeguards. The Equality and Human Rights Commission, advocate for retaining Article 22, arguing that there is a lack of sufficient safeguards to protect data subjects from unfair or discriminatory outcomes.[1030]

[1025] GDPR (n 1) Recital 71.
[1026] ibid Article 22.
[1027] ibid Article 22(1).
[1028] ibid Article 4(4).
[1029] Data Protection and Digital Information Bill (No.2) (n 925) Cl 14.
[1030] Equality and Human Rights Commission, *Data Protection and Digital Information Bill* (House of Lords, 2nd Reading, 15/12/2023) 4-6.

As noted within the literature review, the scope of Article 22 is extremely limited, only applying to decisions 'solely' based on automated processing, providing a noticeable loophole to those decisions which are 'largely' or somewhat based on automated processing.[1031] The literature also highlights the subjectiveness of the phrase 'similarly significant', with concern towards the vagueness of the scope, in addition to the consequential effects on the deployment and use of AI, which may be unnecessarily limited. There has also been academic debate as to whether Article 22 acts as a prohibition, or a right for data subjects to exercise.[1032] On this point, the Article 29 Working Party has argued that the Article is indeed a prohibition,[1033] which this thesis supports. The Article 29 Working Party acknowledge the difficulties in applying the scope of the Article in practice, commenting that it is difficult to be precise as to what would meet the threshold in some applications.[1034] It is essential that AI is subject to rules that are clear, precise, and easily understood by those developing, deploying, and using such systems, to ensure the strongest protections for fundamental rights, in addition to promoting and encouraging a future of ethical AI.

Article 22 also provides exceptions to the right provided, including for contractual purposes, legislative purposes, or based on the data subject's explicit consent.[1035] The first exception included in Article 22(2)(a)[1036] gives rise to concern due to its broad scope, applying in any cases where decisions based solely on automated processing are 'necessary'.[1037] The EP suggest that the ability of AI to outperform human judgement in a given scenario, or a company with a high number of cases to examine could justify automated processing being 'necessary' for use.[1038] However, this exception could lead to abuse, where parties could argue that the use of automated processing is necessary, and therefore do not have to conform to the limitations provided by the right. Concern over the use of subjective terms continues in the second exception provided in Article 22(2)(b), which applies where authorised

[1031] Borgesius (n 4).
[1032] Lee Andrew Bygrave, *The EU General Data Protection Regulation (GDPR) – A Commentary*, (2020, 1st edition, Oxford University Press) 530-532.
[1033] Article 29 Working Party (n 494) 19.
[1034] ibid 22.
[1035] GDPR (n 1) Article 22(2).
[1036] ibid Article 22(2)(a).
[1037] European Parliament (n 247) 61.
[1038] ibid 61.

by law, but that suitable measures must exist to safeguard data subjects. [1039] Included within the GDPR's non-binding recitals, recommendations of such measures include respecting acceptability, accuracy and reliability.[1040] The standard of such 'suitable measures' can be interpreted in various ways, especially between different industries, and therefore needs further clarification to reduce the risk of non-compliance.

The 'suitable measures' standard also exists in Article 22(3),[1041] which applies to both the first and third exceptions. The ICO interprets the standard to suggest that information be given to data subjects on the processing of their data, that accessible and efficient options to challenge decisions and redress are available, and that systems are regularly monitored.[1042] For this to be achieved, companies would need to develop a deeper understanding of the workings of their systems, to ensure that such information and remedies can be offered. The third exception provided in Article 22(2)(c)[1043] requires explicit consent by the data subject, in which the data controller must be able to show that the data subject understands what they are consenting to.[1044] This has also been reinforced within the courts, where it has been ruled that consent is only valid where the data subject is informed about the logic behind the decision-making.[1045] This exception could cause substantial issues and be abused by companies if the means of obtaining consent can be hidden within pages of terms and conditions,[1046] proving inefficient in practice.

Particularly with note to the recent legal challenge against Uber's 'robo-firing' algorithm,[1047] it is evident that the current provisions related to automated decision-making are not sufficient, need further consideration, and could potentially benefit from future reform. These matters are discussed next, with related recommendations proposed to address the concerns that have arisen with Article 22.[1048] In reflection of

[1039] GDPR (n 1) Article 22(2)(b).
[1040] ibid; Article 29 Working Party (n 494) 32.
[1041] GDPR (n 1) Article 22(3).
[1042] Information Commissioner's Office (n 505).
[1043] GDPR (n 1) Article 22(2)(c).
[1044] Article 29 Working Party (n 494) 13.
[1045] Case C-228/21 Request for a preliminary ruling from the Corte suprema di cassazione (Italy) lodged on 8 April 2021 — Ministero dell'Interno, Dipartimento per le Libertà civili e l'Immigrazione — Unità Dublino v CZA (2021) OJ C 217/31.
[1046] Ufert (n 4).
[1047] Wycherley (n 509).
[1048] GDPR (n 1) Article 22.

the limited scope, which only applies to decisions ‘solely’ based on automated processing, the likelihood of data controllers and companies acting to avoid this obligation may arise. In light of this, suggestions are presented to prevent this issue, by way of clarification of the human intervention process relating to automated processing.

Following this, the subjective phrase ‘similarly significant’ is reviewed, highlighting the risk of unnecessary restriction to the development, deployment, and use of AI. The lack of clarity and potential broadness of this scope could risk data controllers, in fear of their systems falling within the scope, opting not to use AI, and consequentially could eventually decrease investment and development. Recommendations are proposed in respect of this, to not only avoid stifling the progression of AI, but to ensure that progression is ethical, and provides effective safeguards to data subjects. Finally, the insufficient exceptions to the right included within Article 22 are explored, with recommendations proposed to prevent data controller misuse, and to further encourage and promote ethical AI.

Decisions based ‘solely’ on automated processing: Recommendations regarding further safeguards to define and investigate ‘human intervention’.

As previously noted, Article 22 provides a noticeable loophole to those decisions which are ‘largely’ or ‘somewhat’ based on automated processing, applying to those only ‘solely’ based.[1049] Although this reinforces the importance for human intervention in relation to high risk AI, the extent of such human intervention may need further assessment. The Article 29 Working Party express that data controllers cannot avoid the Article 22 provisions by fabricating involvement, and that human intervention must have influence on the decision made.[1050]

It should be noted that although England and Wales now have the freedom to depart from EU rules and regulations post-Brexit, the newly introduced Data Protection and Digital Information Bill retains this term of ‘meaningful human involvement’ for the replication of Article 22.[1051] If the same standard was expected as the GDPR, such

[1049] Wachter, Mittelstadt, and Floridi (n 195); Edwards and Veale (n 179).
[1050] Article 29 Working Party (n 494) 21.
[1051] Data Protection and Digital Information Bill (n 925) Cl 14.

human involvement would only qualify in scenarios where the oversight is meaningful, rather than just a token gesture, and that it be carried out by someone with the authority and competence to change the decision.[1052] Under Article 35 of the GDPR, DPIAs should identify and record the degree of any human involvement in the decision-making process, and at what stage this takes place.[1053]

This point of meaningful human intervention has also been noted within the courts. A string of cases have confirmed that several factors are taken into consideration to determine the level of meaningful involvement, including the organisational structure and reporting lines, internal policies and procedures and whether staff have been effectively trained.[1054] In the *Uber* case, it was decided that software used to identify fraudulent activities leading to the deactivation of accounts, were not made 'solely' by automation due to the involvement of a specialised team to initiate investigations to confirm or deny whether activities were indeed suspicious.[1055]

On this point, the CNPD (Portuguese DPA) decision (2021/622)[1056] helps to clarify the extent of meaningful involvement. The decision was made that an automated decision used to detect misconduct during university examinations was 'solely' automated, despite a final decision being made by tutors.[1057] This decision was based on the lack of expertise from the human decision-maker, with the data protection authorities justifying that the decision-maker needs to understand when the automated decision should be followed or not, and without that knowledge, is not enough to state that meaningful human intervention has taken place. The data protection authorities stated that if there were specific guidelines or guiding criteria to help tutors (as human decision-makers) make their decisions, this would have enabled their decision-making to be more meaningful, whilst also ensuring decisions were coherent and transparent.[1058]

[1052] Article 29 Working Party (n 494) 21.
[1053] GDPR (n 1) Article 35.
[1054] Future of Privacy Forum (n 1023) 3.
[1055] *Uber v Uber Drivers (Deactivation Case)* (n 512) [4.19] and [4.2].
[1056] Comissão Nacional de Proteção de Dados, 'Deliberação no. 2021/622' (Portuguese DPA Decision, 2021) < https://gdprhub.eu/index.php?title=CNPD_(Portugal)_-_Delibera%C3%A7%C3%A3o_2021/622> accessed 27th January 2023.
[1057] ibid.
[1058] ibid.

This thesis supports the interpretation that data controllers should not have the ability to avoid Article 22 provisions with the fabrication of human intervention, and that such involvement must provide ‘meaningful oversight’. However, further clarification in the law on the conditions of such oversight and the introduction of consistent monitoring regimes for data controllers may be advantageous to reduce this risk. It must be ensured that the relevant human oversight suffices for the role it intends, and that an objective approach is taken when reviewing automated decisions. Those who regularly work with AI are most likely to fit the Article 29 Working Party’s condition of qualifying human involvement,[1059] that such involvement should be carried out by someone with relevant authority and competence. However, as those individuals regularly work with AI, it must be ensured that they do not become desensitised to AI decisions and impact, which could lead to over-entrustment, and consequentially lead to ineffective intervention.

To reduce this risk, records should be kept of those incidents where human involvement led to a change in decision, and if possible, a ratio of those decisions where human oversight was involved, but no action was taken. This information could then be analysed in consideration of potential improvement of the relevant AI system’s accuracy, to quantify the levels and impact of human intervention and detect any inconsistencies. To go further, it needs to be questioned if such human intervention in automated decisions, whether changes are made or not, should be justified at the time of intervention, so that the interventions made can be reviewed to not only ensure they are sufficient, but to also detect possible human bias towards automated decisions. Details of such reviews could then be sent, collected, and reviewed on a national level, to ensure consistency in the standard of human oversight, and the possible detection of any anomalies. The suggested introduction of an ‘AI Human Oversight Board’ within a new framework could aid in this matter.

Going further than the current stance of DPIAs in the GDPR to identify the degree of involvement and when this takes place, the burden of proof of satisfying human intervention should fall on data controllers who use automated decision-making. For such a burden of proof to exist, ‘human intervention’ would need to be included

[1059] Article 29 Working Party (n 494) 21.

explicitly within Article 22, and also clearly defined. Following this, it can be suggested that:

> Article 22 explicitly defines the term 'solely' as 'decisions made using automated processing, including profiling, that do not satisfy the standard on human intervention'.

Further conditions or requirements defining the standard on human intervention should also be introduced in legislation, to clarify and ensure consistency and effectiveness of human intervention across companies and AI products, and to reduce the risk of company fabrication. For example, the standard of human intervention could be based on five key requirements:

1. *That the relevant individual(s) (perhaps designated 'Intervention Officer(s)') who perform human intervention possess the authority, knowledge, competency, and ability to influence and alter at least, the final decision made by use of automated processing.*

This expands directly on the guidance provided by the Article 29 Working Party who state human involvement should be meaningful and carried out by someone with sufficient authority and competence to change the decision.[1060] Instead of attempting to define the subjective term 'meaningful', the phrase '…with knowledge, competency and ability, can influence and alter at least, the final decision…' is suggested. This would also align with the decision made by the Portuguese DPA (2021/622).[1061] As a standard is being set, these rules would constitute the minimum requirements to satisfy 'human intervention', and for human intervention to be effective, power must exist to alter the final decision made. Emphasis is made that at a minimum, the Intervention Officer(s) must be able to alter the final decision, due to the final decision being most likely to produce output data to a citizen or consumer, and therefore most likely to cause 'legal' or 'similarly significant' effects.

[1060] Article 29 Working Party (n 494) 21.

[1061] Comissão Nacional de Proteção de Dados (n 1062).

2. *That relevant human intervention is made in review of all relevant data, and be more than a simple agreement or disagreement to the final decision. Acts of human intervention and the consequential impact/influence on decisions must be logged and where possible, justified by the Intervention Officer(s).*

For corporations and data controllers to satisfy this requirement, the principles of **transparency**, **explainability**, and **interpretability** in AI systems are imperative. The Intervention Officer(s) should review not only the final decision, but how the technology arrived at such a decision to ensure past decisions were also fair and accurate. Again, this would ensure alignment with the CNPD (Portuguese DPA) decision (2021/622).[1062] Currently, in many machines this may not be possible, whether machines are too complex or not created and developed with the ability to provide interpretable explanations. The introduction of additional 'AI data protection principles' within a new legislative framework could aid in this matter, and in effect make the requirement of justification possible in all scenarios, and therefore an essential condition.

3. *That the logged records of human intervention be reviewed regularly and monitored in-house to ensure effectiveness and consistency in the standard of intervention, and detect possible bias towards automated decisions.*

This requirement would ensure human intervention remains effective for the aims it intends, and would collate data for companies to analyse for the benefit of several factors, such as identifying errors or bias in automated decisions. This review of logged records and justifications could also be reviewed independently of the Intervention Officer(s) to ensure any evidence of human bias towards automated decisions is identified and addressed.

4. *That any unpredictable or unexplainable decisions reviewed are flagged and assessed for potential damage, errors, or bias within automated decision-making.*

[1062] ibid.

This requirement would emphasise the importance of ensuring the ethical progression of AI, and would provide the opportunity for data controllers to identify and correct inaccuracies, or potential bias. In addition to the above requirements, this would also ensure that such intervention is 'meaningful'. Also, this obligation would aid in compliance with the proposed introduction of AI data protection principles within a new framework, which focuses on promoting ethical AI.

5. *The relevant data reviewed and flagged in requirements three and four are collected and sent to the 'AI Human Oversight Board' for review and analysis.*

This requirement is conditional on the introduction of an 'AI Human Oversight Board' or enforcement authority. Currently, under Article 35 of the GDPR,[1063] DPIAs should be completed for 'new technologies', and should specify whether human intervention is involved in their use of automated decision-making systems. The proposed 'AI Human Oversight Board' would in effect, be able to identify the extent to which data controllers perform human intervention, and ensure such data imposed by this requirement is received in the relevant timeframes. This would hold data controllers to account, ensure consistency in the standard of 'human intervention', and also reduce the concern of the ability to fabricate the human intervention process. The data held by the 'AI Human Oversight Board' would also aid the courts, addressing the particular issue brought by the legal claim brought against Uber's 'robo-firing' algorithm, filed in Amsterdam's District Court. [1064]

'Similarly Significant' effects: Recommendations to improve within the Article 22 provision, with the use of an objective standard, the burden of proof placed on data controllers, and demand for strict monitoring requirements.

Currently, the term 'similarly significant' included within Article 22 could give rise to issues concerning the broadness of its scope. In review of the literature and as noted by Privacy International, it is unclear whether the nature of effects is dependent on

[1063] GDPR (n 1) Article 35.
[1064] Wycherley (n 509).

the subjective perception of the data subject or the data controller, or whether an objective standard would be preferred,[1065] reflecting where extra safeguards could be beneficial. This same phrasing is utilised in the Data Protection and Digital Information Bill, and hence, these issues discussed are still relevant for the purposes of England and Wales regulation.[1066] Recital 71 of the GDPR offers limited examples of what could amount to 'similarly significant', including automatic refusal of an online credit application or e-recruiting practices without any human intervention,[1067] but do not add much clarity on the matter.

The Future of Privacy Forum has analysed enforcement decisions made in this area, and suggests that several factors are considered to determine whether a decision fulfils the 'similarly significant' criteria.[1068] These factors indicate that a decision would more likely fulfil this criterion if sensitive data is involved, where decisions have immediate consequences and are long-lasting, where there is a limitation of income or financial loss as a consequence of the decision, and where the data subject can show that the decision was not trivial.[1069] In both of the *Uber* cases, one case had assessed the fraud detection automated process,[1070] and another had assessed the driver and pedestrian matching process,[1071] it was found that the decisions made did not fulfil the standard of 'similarly significant'. The reason behind this was due to the factors outlined above, where the claimants were unable to show that these decisions were long-lasting, and impactful enough to meet the threshold of Article 22,[1072] particularly in the driver and pedestrian matching case, nor were they show the consequences of financial loss from the decision.[1073]

A different decision was made on this point in the *Ola* case, since Ola's algorithm imposed financial penalties on drivers, and hence, it was clearer that the driver's income-making opportunities were affected by the automated decision-making

[1065] Privacy International (n 492) 11.
[1066] Data Protection and Digital Information Bill (n 925) Cl 14.
[1067] GDPR (n 1) Recital 71.
[1068] Future of Privacy Forum (n 1023) 35.
[1069] ibid 35.
[1070] *Uber v Uber Drivers (Deactivation Case)* (n 512).
[1071] *Uber v Uber Drivers (Transparency Case)* (n 1017).
[1072] *Uber v Uber Drivers (Deactivation Case)* (n 512) [4.26]; *Uber v Uber Drivers (Transparency Case)* (n 1017) [4.66-4.67].
[1073] *Uber v Uber Drivers (Transparency Case)* (n 1017) [4.66-4.67].

process.[1074] It is interesting to note that in the *Ola* case, this decision on the applicability of the 'similarly significant' term was taken from the facts, rather than the claimants having the burden of proof to prove this.[1075] A similar standpoint was made by the Garante (Italian DPA), where an algorithm used by Deliveroo took into consideration availability and reliability in offering 'prime-time' weekend shifts.[1076] The Garante held that due to shifts being given, or refused, it was possible to establish the impact of financial loss from the automated decision.[1077]

Outside of the courts, the Article 29 Working Party interprets the phrase 'similarly significant' to mean that the effects of processing must be sufficiently great or important to be worthy of attention.[1078] They add that the decision must have the potential to significantly affect the circumstances, behaviour or choices of the individuals concerned, have a permanent impact on the data subject, and in the most extreme cases, lead to exclusion or discrimination.[1079] This subjective interpretation runs the risk of placing the burden of proof on the data subject, which the first suggestion proposed, contrasts this approach, and clearly places the burden of proof on the data controller. The second suggestion proposed would successfully aid in addressing the discussed concerns. In light of this, to clarify the term 'similarly significant', the following recommendations are proposed:

- That the GDPR place the burden of proof on the data controller, and that an objective standard be used to identify decisions that cause 'similarly significant' effects.

Following the approach taken in the recommendations above for 'human intervention', the burden of proof is placed on the data controller, and contrasts the subjective interpretation made by the Article 29 Working Party.[1080] In practice, a subjective standard when considering 'similarly significant' effects, which takes into account potential vulnerabilities in some data subjects, could be highly restrictive for

[1074] *Ola v Ola Drivers* (n 529), [4.51].
[1075] Future of Privacy Forum (n 1023) 37.
[1076] Garante per la protezione dei dati personali, 'Decision 9685994' (Italian DPA Decision, 2021) < https://gdprhub.eu/index.php?title=Garante_per_la_protezione_dei_dati_personali_(Italy)_-_9685994> accessed 28th January 2023.
[1077] ibid.
[1078] Article 29 Working Party (n 494) 21.
[1079] ibid 21.
[1080] ibid, 21.

AI, and currently, potentially inevitable for data controllers to avoid the current Article 22 provisions. The ICO states several factors that should be taken into consideration in reflection of 'similarly significant' effects, including the extent a decision may affect a data subject's financial circumstances, health, reputation, employment opportunities, behaviour, or choices,[1081] which has been shown by the cases mentioned above.[1082]

The inclusion of 'behaviour' and 'choices' in these examples given by the ICO could have a potentially major unnecessary restriction on AI technology, such as targeted advertising, which heavily relies on automated decision-making.[1083] For this reason, this thesis advocates for the burden of proof to belong to data controllers in these cases where a dispute arises, and a list of factors is drawn up, like those formed by the ICO and within the courts (as pointed out by Future of Privacy Forum),[1084] to assist in determining the applicability of this term. Further processes included within the introduction of a new framework could also allow opportunities for data controllers to challenge their restrictions under Article 22 to the proposed 'AI Human Oversight Board'.

- That the proposed 'AI Human Oversight Board' introduced by a proposed new framework for AI and data protection are given the power to provide monitoring requirements for all AI systems that process personal data, and to resolve Article 22 restriction appeals from data controllers.

This recommendation, although conditional on the introduction of a new framework, could be beneficial for a number of factors. Firstly, it is imperative that data controllers are given sufficient monitoring requirements for all AI systems that process personal data, not only for purposes such as identifying inaccuracies and potential bias in systems, but to also address the specific concerns discussed. If monitoring requirements were introduced, a standard for all AI systems that process personal data would have to be met. For those systems that fall under the provisions provided by Article 22, including those that meet the standard for 'similarly

[1081] Information Commissioner's Office, *Guidance on automated decision-making and profiling* (June 2018) 9.
[1082] Garante per la protezione dei dati personali (n 1082); *Uber v Uber Drivers (Deactivation Case)* (n 512) [4.26] ; *Uber v Uber Drivers (Transparency Case)* (n 1017) [4.66-4.67]; *Ola v Ola Drivers* [2021] (n 529).
[1083] Privacy International (n 492) 11.
[1084] Future of Privacy Forum (n 1023) 35.

significant' suggested above, a stricter set of monitoring requirements could apply to reinforce an ethics-focused approach to AI regulation.

Such monitoring requirements should be independent of acts of 'human intervention' which address individual decisions, whereas examples of monitoring requirements could include regular and extensive checks on systems for data controllers to review and analyse the results on a large scale, to identify and correct any potential inaccuracies or bias, further promoting the progression of ethical AI. As suggested above, an appeal process could be introduced for data controllers who feel an unnecessary restriction is imposed on them in relation to the standard of 'similarly significant', in which appeals could be processed and resolved by the proposed 'AI Human Oversight Board'. This could be done on a case-by-case basis, whereby decisions could be based on using perhaps a test similar to the proportionality test used in human rights legislation,[1085] or a set of requirements for the data controller to prove their automated decision-making should not be considered within the standard set. This 'proof' could, for example, be achieved with the data collated from complying with the suggested 'human intervention' conditions.

The Exceptions to Article 22: Recommendations Relating to Potential Further and Future Safeguards.

In reflection of the challenges posed by the Article 22 provision, concerns are raised regarding the exceptions to the right, namely for contractual purposes, for legislative purposes, or with explicit consent of the data subject.[1086] Further safeguards are needed to ensure these exceptions are not overused or misused by data controllers, and to encourage the progression of ethical AI in cases where such exceptions do apply. It should be noted that similar exceptions are included within the Data Protection and Digital Information Bill.[1087] In light of this, three recommendations are proposed in respect of each exception provided within the regulation, to add clarity to the law, and demonstrate how the introduction of AI data protection principles within a new framework could aid in these matters.

[1085] Human Rights Act 1998.
[1086] GDPR (n 1) Article 22(2).
[1087] Data Protection and Digital Information Bill (n 925) Cl 14.

- The exception given for contractual purposes: Clarity and review of the term ‘necessary’ to ensure the reduction in the likelihood of data controller overuse or misuse.

Under Article 22(2)(a), the exception is provided in matters where it “*is necessary for entering into, or performance of, a contract between the data subject and a data controller*”.[1088] The use of the term ‘necessary’ here needs clarification, and could lead to problematic arguments on the use of automated decision-making. The Future of Privacy Forum highlights the area for recruitment that could lead to issues for this exception;[1089] where AI practices in this field have already led to severe ethical concerns.[1090] To reflect this, the forum refers to the example of Germany and the use of AI to analyse CVs and cover letters to determine whether an applicant continues or not in the recruitment process.[1091] The guidance provided suggests that the use of such a system would not pass the threshold for 'necessary'.[1092] Although this guidance is helpful for this one example, broader guidance is needed.

The standard of ‘necessary’ could follow the approach of the above recommendations by placing the burden of proof on data controllers to justify such use is ‘necessary’ based on several set factors, before such use of systems can be used within relevant contracts. The suggested introduction of an ‘AI Human Oversight Board’ within a new framework for AI and data protection could aid in reviewing such justifications provided by data controllers, and grant or reject consequential use of the exception. As suggested by the EP, relevant factors which could be considered include the capability for AI to outperform human judgement in the given scenario, or the amount of data to process or analyse.[1093] If refurbished data protection principles for AI were added into a new framework, this could help determine the scope of ‘necessary’, ensuring that this exception is relied on only in ethical circumstances.

[1088] GDPR (n 1) Article 22(2)(a).
[1089] Future of Privacy Forum (n 1023) 9.
[1090] Zhisheng Chen, ‘Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices [2023] 10 *Humanities and Social Sciences Communications* 567; Dastin (n 331).
[1091] Future of Privacy Forum (n 1023) 9.
[1092] ibid 9.
[1093] European Parliament (n 247) 61.

- The exception given for legislative measures: An obligation to comply with binding AI data protection principles introduced in a new framework focused on promoting ethical AI.

Under Article 22(2)(b), the exception provides for circumstances authorised by law, where suitable measures to safeguard the data subject's rights, freedoms and legitimate interests are implemented.[1094] Such 'suitable measures' are said to include at least the right to obtain human intervention on the part of the controller, and to challenge and contest the related decision.[1095] The ICO interprets that such measures should include the data subject being given information on the processing completed by the data controller, that accessible and efficient options are available to challenge decisions, and that systems are regularly monitored.[1096] In a 2021 decision, the Slovakian Constitutional Court suggested that 'suitable measures' need to go further than this, particularly in circumstances where state agencies use automated assessments.[1097] The court stated that additional mechanisms should ensure that:

- databases used to underpin automation are up to date, reliable and non-discriminatory,
- that individuals are aware of the existence, scope, and impact of the automation,
- that systems are quality-checked for errors before and during the system is used, and
- that redress rights are readily available.[1098]

The Article 29 Working Party suggests that suitable measures should also include regular monitoring to reduce bias, the introduction of processes to prevent errors, inaccuracies and discrimination, which aligns with the above decision by the Slovakian court, and the right to obtain an explanation of the decision reached.[1099] This has been consistent in Europe, where the lack of accuracy was cited as the basis of unlawful use of automated decision-making by the Garante (Italian DPA).[1100]

[1094] GDPR (n 1) Article 22(2)(b).
[1095] ibid Article 22(3).
[1096] Information Commissioner's Office (n 505).
[1097] *eKasa System Case* (*Ústavného súdu Slovenskej republiky)* (2021) Case 492/2021 Z. z.
[1098] ibid [132-135] and [137-138].
[1099] Article 29 Working Party (n 494) 32.
[1100] Garante per la protezione dei dati personali (n 1088).

This decision followed the earlier case of *Foodinho*, and highlights the approach from Italy in forcing the upkeep of such suitable measures.[1101]

Currently, however, the standard of 'suitable measures' in law remains unclear, and would substantially benefit from binding AI data protection principles, especially given England and Wales's current freedom to depart from the GDPR rules post-Brexit. Data protection rules could then reference the standard of 'suitable measures' to the new framework, providing clarity and cohesion between AI regulation and data protection. AI data protection principles could also aid in ensuring a human rights centred approach to AI, and could include as discussed in Chapter Three, **transparency**, **explainability** and **interpretability**, **lawfulness**, **non-discrimination, accuracy,** and **fairness** to set, consistent and appropriate standards.

Data controllers would also be under obligations to comply with the monitoring requirements suggested above, where a higher standard of monitoring would apply for decisions that would make legal, or similarly significant effects (using the redefined interpretation of this term). Regardless of whether Article 22 is triggered or not, a higher and stricter standard for applying all principles could exist. This would ensure that any automated decision-making that falls within the exceptions provided by Article 22 would still have to comply with strict standards and binding rules, and therefore provide sufficient safeguards to data subjects, and clearer rules for data controllers.

- The exception given based on explicit consent from the data subject: Additional safeguards to ensure the reduction in the likelihood of data controller overuse or misuse.

Under Article 22(2)(c), the exception is given based on the data subject's explicit consent.[1102] 'Explicit consent' is not defined in the GDPR, but is interpreted by the ICO to involve a specific, informed and unambiguous indication of the individual's

[1101] ibid.

[1102] GDPR (n 1) Article 22(2)(c).

wishes, and that it be affirmed in a clear statement.[1103] Concern has stemmed from fear of data controllers being able to hide such consent within pages of terms and conditions, which is rarely read and understood in full by the data subject before agreeing, deeming such an exception inefficient for the job it intends.

The EDPB states that consent needs to be expressed formally, and suggests that a written statement of consent would be sufficient, ideally signed by the data subject for evidence purposes,[1104] which would reduce the above concern. The ICO in addition to this states that data controllers must be able to show that the data subject has understood exactly what they are consenting to.[1105] This is difficult to achieve in practice, especially concerning AI technologies which are complex, and may not be fully understandable by the reasonable layman. Italy's Supreme Court has also reinforced the importance of this point, where it was held that consent is invalid where there is inadequate information on the logic behind automated decisions,[1106] which reflects the responsibility of data controllers when relying on this exception.

In consideration of online services, a potential solution to this could be one similar to the 'cookie banner' used by data controllers to acquire cookie consent imposed by the GDPR and ePrivacy Directive.[1107] When such cookie banners were introduced, major awareness arose in society and the media to the changes on websites,[1108] and consequentially made the public more aware of cookies, and their purposes. This could also be the case for automated decision-making, where society is made more aware, consequentially having the potential to increase public trust. Such a pop-up could arise before related decision-making takes place, not allowing the data subject to process further without consent. It is important to note the obvious here, in terms of laymen automatically consenting to the cookie banner without due reading of its contents.

[1103] Information Commissioner's Office, *Guidance on Lawful basis for processing: Consent'* (March 2018) 3.

[1104] European Data Protection Board, *Guidelines 05/2020 on consent under Regulation 2016/679* (2020) 20-21.

[1105] Information Commissioner's Office (n 1109) 28.

[1106] *Algorithm Transparency Case (Garante per la Protezione dei Dati Personali Associazione Mevaluate Onlus)* (2021) Case 14381/2021.

[1107] GDPR (n 1); Council Directive (EC) 2009/136 concerning the processing of personal data and the protection of privacy in the electronic communications sector OJ L 337.

[1108] Jack Schofield, 'What should I do about all the GDPR pop-ups on websites?' *The Guardian* (5th July 2018); Emily Stewart, 'Why every website wants you to accept its cookies' (Vox Blog, 10th December 2019) <https://www.vox.com/recode/2019/12/10/18656519/what-are-cookies-website-tracking-gdpr-privacy> accessed 8th March 2021.

Consent should also be based on two statements; firstly, that the data subject consents to automated decision-making being used, and that such decision-making does not include human intervention, and secondly, that the data subject consents to the risk of unpredictable decisions made by automated decision-making, with options of redress readily available. Such a pop-up could also include links to educate users such as ‘What is automated decision-making and how is it used?’ or ‘How are my rights affected?’ and would allow data controllers to state compliance to the suggested monitoring requirements and the proposed AI data protection principles, in addition to redress mechanisms. However, like the Cookie banner, the risk of data subjects consenting without reading should be acknowledged.

4.4 A Proposed New Framework for AI

As noted previously, to sufficiently tackle the complex and novel challenges brought by AI, new legislation focused on addressing these matters is essential. As proposed throughout this thesis, a new framework for AI could complement, and work in collaboration with existing legislation, such as the GDPR, to add clarity for data controllers, and strengthen protections for data subjects. As noted by Borgesius and Wachter et al, several algorithmic decision-making processes fall outside of the scope of the GDPR,[1109] highlighting the need for new regulation to ensure AI does not undermine human rights. A new legislative framework could aid in filling such gaps in existing legislation, and binding legislative obligations would also ensure AI systems are held to a consistent standard, proving significantly more useful than soft law approaches. In addition to and for clarification of the Article 22 reform suggestions discussed above, a new legislative framework for AI could aid and complement these recommendations in the following ways:

4.4.1 Establishment of an ‘AI Human Oversight Board’

When new AI legislation is introduced, it needs to have a human rights centred approach, ensuring not only sufficient protections and safeguards for data subjects, but also to promote ethical AI. For such legislation to be efficient, an enforcement board could aid in providing clarity of provisions, to ensure all parties understand and

[1109] Borgesius (n 4); Wachter, Mittelstadt, and Floridi (n 195).

can comply with their obligations. As suggested above, the introduction of an AI Human Oversight Board could aid in the application of the 'human intervention' and 'similarly significant' standards, and consequentially enforcement of the Article 22 provisions. A board could also be given power in a new framework to publish redress and monitoring requirements, and as suggested above, with a higher and stricter standard for any automated decision-making that produces legal or similarly significant effects. The application and scope of the exceptions provided by Article 22 of the GDPR could also be clarified with the support of an AI Human Oversight Board, such as the standard of 'necessary' within the contractual exception. With the establishment of such a board, opportunities, and prospects for those in the AI industry would also arise, giving the opportunity to raise public awareness and education on AI.

4.4.2 Introduction of Refocused AI Data Protection Principles

As noted in Chapter Three, the introduction of re-focused data protection principles specifically for AI systems could work in collaboration with the current data protection principles provided by the GDPR,[1110] ensuring a stronger and more sufficient standard for AI systems. As discussed, such principles could include **transparency**, **explainability** and **interpretability**, **lawfulness**, **non-discrimination, accuracy,** and **fairness**, and form the foundations of future legislation. Specified AI data protection principles could also aid in several other matters concerning AI regulation and its alignment to the GDPR, such as providing solutions to the extensive debate on the 'right to explanation' within Articles 13-15.[1111] Such principles could also aid in clarifying several points within the law, and would align with the principles of accessibility and foreseeability set out by the courts.[1112]

With such principles in place for relevant AI systems, the recommended 'AI Human Oversight Board', in review of evidence from data controllers on 'human intervention' and redress and monitoring obligations, could identify compliance issues or difficulties with automated systems. This would allow for easy identification of the need for further protections and safeguards, and adjustments could be

[1110] GDPR (n 1) Article 5.
[1111] ibid Articles 13-15.
[1112] Joined cases C-203/15 and C-698/15 (n 838) AG Opinion [137-154]; *Szabó and Vissy v. Hungary* (n 979).

recommended where necessary. This could potentially achieve the eventual aim of removing the restrictive nature of the right provided within Article 22 of the GDPR,[1113] which although currently necessary, the clear limit on the progression and development of AI systems must be noted.

4.5 Conclusion

This chapter has critically examined the limitations of the GDPR in addressing the unique challenges posed by AI. The suggestions proposed relating to the adjustment of the GDPR provisions on automated decision-making and profiling aim to resolve ambiguities, enhancing the applicability of the GDPR to AI, whilst also promoting ethical practices. These suggestions are intended not only to clarify the scope of the provisions, but also to ensure more effective safeguards for individuals affected by AI decision-making. Through establishing clearer obligations for data controllers using AI, the proposed suggestions aim to reduce legal uncertainty, and are to be read in combination with the proposed new framework for AI, working collaboratively to effectively promote ethical AI. Through providing strong safeguards to individuals, and guidance to developers, understanding of compliance would be eased.

The thesis advocates for the GDPR to be used as a basis and foundation for the proposed future framework for AI, to ensure not only that strong protection of data protection still exists, but also that the provisions explicitly address and are focused on the complexities of AI, providing developers with clear and workable guidelines for compliance. Such a framework should build upon the GDPR's strengths, whilst addressing and filling its gaps, particularly in reference to the requirements for human intervention and the right to an explanation. In light of the GDPR data protection principles, although AI systems need to be subject to these, additional principles or safeguards are needed to ensure alignment between AI and data protection. Through filling the gaps within the GDPR, AI can develop in a manner that upholds and safeguards human rights whilst also fostering ethical innovation.

In light of the difficulties within Articles 13-15 and Article 22, the suggestions recommended aim to provide clarity on the scope of the provisions, to make them

[1113] GDPR (n 1) Article 22.

easier to apply to the complexities of AI technology, and to ensure ethical AI development. The broader analysis and recommendations proposed in Chapter Three are interlinked to the suggestions proposed in this chapter. Together, the proposals in these chapters form the basis of a holistic approach to AI regulation, which is further examined in Chapter Five through a review of current regulatory attempts and ongoing legislative developments.

The thesis concludes by combining the suggestions, to offer a recommended structure, and suitable provisions to be included in a new legislative framework for AI that not only addresses the complexities of AI, but also places human rights at its core. By combining the suggested reforms to the GDPR with new, AI specific provisions, a balanced and effective framework can be established. The final chapter will present these recommendations, proposing a framework which answers the overarching question that underpins this research; ‘to what extent should AI be regulated to ensure an ethical framework centred on human rights?’.

Chapter 5: AI and Regulation

RQ3: *What limitations exist in recent attempts to regulate AI made by the EU and UK, and in reflection, what solutions can be proposed to strengthen safeguards whilst encouraging ethical innovation?*

5.1 Introduction

This chapter aims to use a doctrinal approach to build on the discussions in Chapter Three and Four by evaluating the recent regulatory developments for AI in the EU, with a focus on the AI Act,[1114] and the UK's evolving proposals.[1115] In previous chapters, the thesis explored the ethical challenges and legal complexities posed by AI, including the limitations of the GDPR when applied to AI. This chapter aims to assess how proposed regulatory frameworks attempt to address the challenges noted, beginning with the considerations that have been raised in response to the recently approved EU regulation. The chapter also intends to assess the UK proposals[1116] towards regulation, making a comparison with the EU, and identifying the parts to be commended and those that may need to be re-worked.

By drawing on both the EU and UK approaches, the chapter will make use of a comparative approach, with predominantly a macro comparative element, to examine prominent proposals from other jurisdictions,[1117] but also some micro comparative elements, to reflect on the issues that need to be at the forefront of AI regulation, which were introduced in Chapter Three in regard to ethical considerations and Chapter Four on the role of data protection in AI governance. The thesis intends to conclude in Chapter Six by addressing the latter half of RQ3, connecting the proposed recommendations in Chapters Three and Four to the current regulatory landscape explored within this Chapter.

[1114] Provisional Agreement for the AI Act (n 103).
[1115] Department for Science, Innovation and Technology (n 99).
[1116] ibid.
[1117] White House, *Blueprint for an AI Bill of Rights* (October 2022); White House, *Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence* (October 2023); An Act to amend Sections 12930 and 14203 of the Government Code,and to amend Section 156 of 90.5 of, and to add Part 5.6 (commencing with Section 1520) to Division 2 of, the Labor Code, relating to employment (Assembly Bill No. 1651) 2022; Digital Charter Implementation Act 2022 C-27 (44th Parliament, 1st Session, 2021-2022); Consumer Privacy Protection Act 2022 Bill C-27 (44th Parliament, 1st Session, 2021-2022); Artificial Intelligence Bill (translated) (Federal Senate Bill, No. 21 of 2020).

The chapter will begin with an overview of the EC's White Paper on AI,[1118] introduced back in 2020, which was the first mainstream proposal for AI regulation at the time. The chapter will build on this overview by introducing the AI Act, first proposed in 2021,[1119] with intense trilogue debates taking place during 2023, and approval granted in December 2023, with publication expected in 2024.[1120] It is important to note that the latest draft of the Act has not yet been officially published, so for the purposes of this discussion, the latest draft (which was endorsed in the COREPER (Council of Ministers' Permanent Representatives Committee) meeting in February 2024)[1121] will be included as part of the discussion. Although there may be minor changes to the final draft (due for plenary adoption in April 2024), the current agreed text is expected to be the final compromise text, so any potential changes would be minimal, making little impact on the analysis provided within this Chapter.

Through this discussion, the thesis will draw out articles that should be commended, and those that have raised concerns from academics, data protection authorities, and human rights campaign groups. The thesis will reflect on the likelihood of compliance with the GDPR when the AI Act is enforced, and highlight where issues may arise. The chapter will include a case study based on the *R(Bridges)* case,[1122] to assess how AI technology would comply with the Act, and reveal whether the proposals are sufficient for everyday and real-world practice. The chapter will also review the EU's proposed AI Liability Directive of 2022,[1123] assessing whether the proposal would satisfy the issues highlighted in the discussion of liability discussed in Chapter Three.

Following this, the chapter will assess the recent UK Government's strategy, through examination of the published AI Policy and White Paper,[1124] and the newly formed Artificial Intelligence (Regulation) Bill,[1125] with comparisons made to the EU to

[1118] European Commission (n 112).
[1119] European Commission (n 117).
[1120] European Parliament, 'Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI' (European Parliament Press Release, 9th December 2023) < https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai > accessed 15th December 2023.
[1121] Provisional Agreement for the AI Act (n 103).
[1122] R(Bridges) v CC of South Wales Police (n 49).
[1123] Proposed AI Liability Directive (n 104).
[1124] Department for Science, Innovation and Technology (n 99); UK Government, *Policy Paper- Establishing a pro-innovation approach to regulating AI* (July 2022).
[1125] Artificial Intelligence (Regulation) Bill (n 101).

highlight the similarities and differences to the regulatory approaches, and the criticisms raised. Subsequently, the chapter will review other recent prominent proposals, namely proposals from the US, California, Canada, and Brazil, to comment on whether inspiration can be taken from outside the Europe-sphere. To conclude the chapter, the thesis will tie together compliments to sections of proposals and highlight recommendations to strengthen protection and safeguards further. This analysis will set the stage for the final chapter, where the thesis will combine the findings from across the chapters to propose a comprehensive framework for AI regulation, which aligns with the central research question of the thesis.

5.2 AI Regulation in the EU

The EU has made clear for several years that they intend to be at the forefront of AI regulation, a stance that was similarly taken during the drafting of the data protection regulation, which has since been used as an inspiration worldwide. The EU has lived up to this stance, with the early introduction of the White Paper on AI in 2020.[1126] The White Paper highlights the initial aims of the EU in seizing the opportunities offered by AI, promoting Europe's innovation capacity and supporting the development of ethical and trustworthy AI across the economy, with the idea of AI working effectively for the people and being a force for good.[1127] The White Paper, although commended for being one of the first bold moves towards regulation, reflected the real complexities of detailing a successful AI framework and has since been heavily criticised, interestingly for reasons that were also brought forward to the AI Act, including the foundations of the risk-based approach, and adaptions made from the draft version of the document.[1128]

There have also been rising concerns related to the funding of regulation, and the associated costs to businesses and consumers. The Foundation for European Progressive Studies comments that the focus on regulation and enforcement could lead to Europe losing its competitive advantage in digital governance, and overall

[1126] European Commission (n 112)
[1127] ibid 25.
[1128] MacCarthy and Propp (n 671); Virginia Dignum, Catelijne Muller and Andreas Theodorou, *Final Analysis of the EU Whitepaper on AI* (ALLAI, June 2020) 7; William Crumpler, 'Europe's Strategy for AI Regulation' (CSIS Blog, 21st February 2020) <https://www.csis.org/blogs/strategic-technologies-blog/europes-strategy-ai-regulation> accessed 20th July 2023.

advancements and innovation of technology, whilst other jurisdictions are focused on investing in digital capacity.[1129] It is estimated, according to the Center for Data Innovation, that the proposed AI Act when enforced will cost the European economy up to €31 billion over the next five years, in addition to reducing technical investments by 20%, stifling advancements in innovation, and profits of high risk AI companies dropping by 40% to comply with the regulation.[1130] Due to these expected compliance costs, the effects on small and medium enterprises and start-up companies would be larger, deterring smaller companies from engaging in the technology, and hence, slowing down the digitisation of the economy.[1131] This would effectively give the market to the already prominent 'big tech' companies.[1132] A similar negative effect on innovation was seen after the implementation of the GDPR, where it was estimated that one-third of mobile apps were removed from e-stores due to the increased costs of compliance.[1133]

Regardless of these factors, the EU has continued efforts to develop regulation, leading to the well-reported introduction of the AI Act proposal in April of 2021, which the EU highlighted was the first-ever proposal of its kind, positioning Europe to play a leading global role in the regulation of AI.[1134] The AI Act applies horizontally to all sectors and industries, and works on a risk-based approach, labelling different AI technologies as posing differing risks to society, and therefore having to conform to the rules related to that risk factor. The theory behind this approach is to ensure users and those affected by AI can trust the machines that are being used, strengthening the EU's ability to compete on a global level.[1135] The proposed regulation, which has recently gained approval from the EU institutions, is a development on the feedback received from the White Paper, and plays a part in the

[1129] Andrea Renda, *Beyond the Brussels Effect* (Policy Brief, Foundation for European Progressive Studies, March 2022) 12.
[1130] Center for Data Innovation, *How Much Will the Artificial Intelligence Act Cost Europe?* (July 2021) 3.
[1131] Luca Bertuzzi, 'Making the AI Act work for SMEs: The EU tries to square the circle' (Euractiv Blog, 25th November 2022) <https://www.euractiv.com/section/digital/news/making-the-ai-act-work-for-smes-the-eu-tries-to-square-the-circle/> accessed 22nd July 2023.
[1132] European Digital SME Alliance, *Digital SME reply to the AI Act consultation* (6th August 2021) 1-3.
[1133] Rebecca Janßen, Reinhold Kesler, Michael E Kummer and Joel Waldfogel, 'GDPR and the Lost Generation of Innovative Apps' (NBER Working Paper No. w30028, July 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4104014> accessed 1st August 2023.
[1134] European Commission (n 117).
[1135] ibid 1.

EU's overall Coordinated Plan on AI, which aims to be a big step in ensuring EU global leadership in trustworthy AI.[1136]

The EC's Coordinated Plan on AI,[1137] like the AI Act proposal, was first published in April of 2021, and sets out the intended strategy to accelerate investments in AI to drive digital solutions. This includes acting on AI strategies with timely implementation to ensure the full benefits are reaped, and to align AI policy to address global challenges,[1138] which conforms with the earlier aims set out in the Communication on AI for Europe in 2018,[1139] and the 2020 White Paper on AI.[1140] The plan makes clear its aims to seize the opportunities and benefits brought by AI, and that the focus of the approach is based on ensuring AI is human-centric, trustworthy, secure, sustainable, and reflects core European values.[1141] When the AI Act is enacted, similar to the GDPR, and the recently passed Digital Services Act and Digital Markets Act,[1142] the EC would be predominantly influential in governing online platforms, surpassing any other democratic government.[1143]

The AI Act's scope is similar to the GDPR,[1144] applying to providers and deployers of AI systems that are available in the EU, or where the output produced by the system impacts those within the EU, regardless of where the provider or deployer itself is located,[1145] and hence, having an extraterritorial effect. 'Providers' refers to developers of AI, in addition to individuals or bodies who put products on the market under their name or trademark.[1146] The term 'deployer' refers to individuals or bodies

[1136] European Commission, *Annexes to the 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- Fostering a European approach to Artificial Intelligence, COM (2021) 2.
[1137] ibid.
[1138] ibid 2.
[1139] European Commission, *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- Artificial Intelligence for Europe,* COM (2018).
[1140] European Commission (n 112).
[1141] European Commission (n 1141) 4.
[1142] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277; Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) [2022] OJ L 265.
[1143] Alex Engler, 'The EU AI Act will have global impact, but a limited Brussels Effect' (Brookings Blog, 8th June 2022) <https://www.brookings.edu/articles/the-eu-ai-act-will-have-global-impact-but-a-limited-brussels-effect/> accessed 5th August 2023.
[1144] GDPR (n 1).
[1145] Provisional Agreement for the AI Act (n 103) Article 2(1).
[1146] ibid Article 3(2).

that use AI under their authority to carry out tasks.[1147] The risk-based approach that underpins the Act is based on four separate risk levels, ranging from a low (minimal and limited) risk, to high risk,[1148] and unacceptable risk;[1149] those labelled as unacceptable are outright banned.[1150] At first glance, and considering the risks posed to human rights by AI highlighted in Chapter Three, the notion of banning systems that pose unacceptable risks may bring a sense of relief.

However, realistically, as detailed later in this Chapter, the listed systems contain several broad and vague exceptions,[1151] opening the risk of loopholes existing for such ‘unacceptable’ technology, raising questions about the practical implications of this approach. One issue is the broad and at times, vague definition of what constitutes ‘unacceptable’ risk, in addition to the broad exemptions given to the systems. These exemptions risk disrupting the strength and significance of the prohibition of systems, undermining the very protection the regulation intends to provide.

In light of this, it is worth questioning whether the risk-based framework is truly sufficient to prevent the deployment of harmful systems. The thesis argues that whilst the EU’s risk-based framework has commendable intentions, more specific and enforceable standards are required. This leads to a broader question of whether a more precautionary approach would be more effective in regulating dangerous AI. By implementing a future framework which puts more focus on the potential impact on society and less on categorising risk, the precautionary model offers a more adaptive and clearer framework, mitigating risks before they manifest. Such an approach aligns with the overall thesis, which advocates for a regulatory framework which centres on human rights and ethical considerations at the core of AI governance, ensuring that safeguards are not undermined by vague exceptions or inconsistencies.

[1147] ibid Article 3(4).
[1148] ibid Article 6.
[1149] ibid Article 5.
[1150] ibid Article 5(1).
[1151] ibid Article 5(1)(d)(i-iii).

Those systems identified as high risk are the focus of the EU's regulation, and most Articles focus on setting out rules to regulate such systems. The proposal does not provide a clear definition of 'high risk', but does set out two conditions that both need to be fulfilled to amount to the categorisation:

> "*(a) the AI system is intended to be used as a safety component of a product, or is itself a product, covered by the Union harmonisation legislation listed in Annex II;*
>
> *(b) the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or putting into service of that product pursuant to the Union harmonisation legislation listed in Annex II*".[1152]

For context, Annex II contains a list of legislation that already exists that covers the safety of products and components, such as medical devices and machinery.[1153] Article 6 also specifies that in addition to those systems that satisfy the two above elements, those listed in Annex III are also to be considered high risk,[1154] which include examples such as educational and vocational training systems and employment management systems.[1155] The list within Annex III intends to be updated annually,[1156] in circumstances that satisfy the conditions set out in Article 7.[1157]

The systems categorised as limited and minimal risk have little mention in the Act, and hence, are only subject to minimal rules depending on the purpose and use of systems. Article 69 of the proposal[1158] supports the ethos of the regulation by stating that the EC and MS should encourage and facilitate the creation of codes of conduct for those systems not considered high risk, in which they can voluntarily conform to ensure compliance with other parts of the regulation. The regulation can be viewed as revolutionary for taking the first steps at regulating AI as a whole, which has been sought after for several years from several sectors, but it has received bounds of criticism that will be highlighted in the next section of this chapter.

[1152] ibid Article 6(1).
[1153] ibid Annex II.
[1154] ibid Article 6(2).
[1155] ibid Annex III(3).
[1156] ibid, Article 73.
[1157] ibid, Article 7.
[1158] ibid, Article 69.

Developing the EU's initiative to be world-leading in AI regulation, the proposed AI Liability Directive was published in September of 2022 [1159] tackling the challenges identified in the Report on AI Liability,[1160] a report released accompanying the White Paper on AI.[1161] The initial Report on Liability in 2020[1162] emphasised the EU's objective of striving for the safety and trustworthiness of AI, and that a clear liability framework is of particular importance to ensure these objectives can be achieved, in addition to ensuring a sufficient level of consumer protection and legal certainty.[1163] Building on this report,[1164] the AI Liability Directive[1165] considers the comments received from the consultation response to the 2020 report,[1166] and aims to clarify the rules on safety and liability. The Directive highlights that current liability rules are inadequate at handling claims that involve AI-enabled products, rendering victims unable to access sufficient redress, and potentially deterred from making claims altogether.[1167]

The Directive intends to complement the AI Act, and seeks to achieve similar aims by contributing to the enforcement mechanisms related to the systems categorised as high risk, as a failure to comply with the rules within the Act can play a part in helping identify fault.[1168] The Directive also intends to set a harmonised standard to avoid the current fragmentation of liability initiatives in MS, to ensure that AI products can be used in a consistently safe manner within and between nations.[1169] To achieve its aims, the proposal takes a staged method, firstly with a minimally invasive approach through the introduction of burden of proof measures to address the particular problems identified relevant to AI, and secondly, assessing whether more stringent measures are needed to protect society through evaluating the effects of the first stage.[1170]

[1159] Proposed AI Liability Directive (n 104).
[1160] European Commission, *Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee- Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics,* COM (2020)
[1161] European Commission (n 112).
[1162] European Commission (n 1163).
[1163] ibid 1.
[1164] ibid.
[1165] Proposed AI Liability Directive (n 104).
[1166] European Commission (n 1163).
[1167] Proposed AI Liability Directive (n 104).
[1168] ibid 3.
[1169] ibid 5.
[1170] ibid 6.

Moving forward, this chapter intends to focus on the two most recent regulations from the EU, the AI Act, and the proposed AI Liability Directive, highlighting the criticism raised within the literature, whilst also commending sections of the documents that seem likely to achieve the overall aims of promoting ethical AI. Through this discussion, the thesis will include some example case studies to test the practicalities of the frameworks as they would intend to work in the 'real world'.

5.2.1 The Artificial Intelligence Act

A key issue in the regulation of AI has been producing a standardised definition, accepted by all sectors relevant to the AI industry. In the AI Act proposal, a lengthy definition within Article 3(1) was originally given, defining an AI system as "*software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with*".[1171] This definition was subject to intense deliberation during the trilogue debates and has since been changed. This original definition was a much more developed offering than one of the many definitions given in the White Paper, one of which defines AI simply as a "*collection of technologies that combine data, algorithms and computing power*".[1172]

However, the first AI Act definition was not as detailed as the one given in the EC's Communication on AI back in 2018.[1173] This more detailed definition was also supported by the HLEG on AI, responsible for the publication of the Ethics Guidelines for Trustworthy AI[1174] in 2019, and defined AI as "*systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be*

[1171] European Commission (n 117) Article 3(1).
[1172] European Commission (n 112) 2.
[1173] European Commission (n 1144) 1.
[1174] High-Level Expert Group on Artificial Intelligence (n 100).

embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)“.[1175]

The original definition in the AI Act was a standardisation of all three above, with obvious reference to other parts of the regulation to extend on and clarify the definition, a definition the EU called itself a “*single, future-proof definition of AI”*.[1176] However, as noted, in the final draft of the AI Act, the definition has since changed, making the original not future-proof at all. The definition of AI now stands as the following: “*a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content recommendations, or decisions that can influence physical or virtual environments”.* The definition combines the autonomy and adaptiveness factors and could be said to still be quite broad, however, it is a clear improvement on the prior suggestions.

It is important to have a standardised definition for future legislation, but Schuett highlights the need for preciseness when it comes to legal definitions, and as AI is an umbrella term which is used for a broad range of systems, preciseness is not always possible.[1177] For this reason, it is important that the definition is clarified and updated regularly, to ensure that it can advance into the future and remain aligned to the technology it defines. In support of this view, the AI Act proposal stated that the definition should be complemented with a list of techniques and approaches used for its development, which is regularly updated through the use and introduction of delegated legislation.[1178] This would ensure clarification to the definition, and make sure it is sufficient not only for the present, but also for the future.

The Czech Presidency of the Council of the EU had reportedly shared concerns related to the scope of the original definition provided in that it is too broad and

[1175] ibid 1.
[1176] European Commission (n 117) 3.
[1177] Jonas Schuett, ‘Defining the scope of AI regulations’ [2023] 15(1) *Law, Innovation and Technology* 60.
[1178] European Commission (n 117) 18,

ambiguous, and the extent to which it can be adapted in the future.[1179] This was a view shared by many, with a body representing the UK Tech Industry, Tech UK, stating that the original definition went beyond the consideration of intelligence, and the Act would also cover general software and computer programs that would not align with the traditional understanding of AI.[1180] This view is echoed by Huawei and IBM, who share the concern that the Act may govern systems that are not intended to be included.[1181] In light of the newly introduced definition in the final version of the AI Act, the scope of the definition has been refined, with the focus made on inputs and outputs, which reduces the concern of unwanted and unintended products falling within the definition.

The 'new and improved' definition reveals a clear comparison with the updated OECD definition of AI: "*a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment*".[1182] In light of this, MEP Maydell highlights that the attempts to align with the OECD definition should be commended, as it shows consistency in setting a global standard.[1183] The rationale behind the general broadness of AI definitions is that more focus needs to be put on categorising systems on the intended use of AI in a narrow manner, rather than focusing on defining an AI system in itself more precisely.[1184] The broadness of the scope has

[1179] Luca Bertuzzi, 'Czech Presidency sets out path for AI Act discussions' (Euractiv Blog, 22nd June 2022, updated 28th June 2022) <https://www.euractiv.com/section/digital/news/czech-presidency-sets-out-path-for-ai-act-discussions/> accessed 18th December 2022.

[1180] TechUK, *techUK response to the Commission's proposed Artificial Intelligence Act* (Feedback Reference: F2665579, August 2021) 1.

[1181] Huawei Technologies, *Huawei response on the European Commission's Proposal for a Regulation of the European Parliament and of the Council Laying Down the Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (Feedback Reference: F2665442, August 2021) 3; IBM, *IBM Submission to the European Commission on the Draft Artificial Intelligence Act* (Feedback Reference: F2665615, August 2021) 2.

[1182] OECD (n 626) 7.

[1183] European Parliament, *Draft Opinion of the Committee on Industry, Research and Energy* (Rapporteur for opinion: Eva Maydell, March 2022) 5.

[1184] Luciano Floridi, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander and Yuni Wen, *capAI- A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act* (Version 1.0, March 2022) 9.

also been said to be necessary to ensure that exceptions are not made for general-purpose systems, and that if needed, adaptions could be made in the future.[1185]

The foundation of the regulation as stated in the above section works on a risk-based approach, labelling systems according to the risk posed to society; the highest risk being an unacceptable risk, then high risk systems, followed by low risk (limited or minimal risk) systems. The risk-based approach is not a new methodology for legislation, and was also used as a premise for the GDPR. Within the AI Act proposal, this methodology is commended as a proportionate approach, that ensures there are no unnecessary restrictions to trade, whilst also ensuring intervention is possible to systems that pose concerns for society.[1186] The AI Act itself focuses predominantly on the rules relating to high risk systems, with low risk systems having very little reference and only a few Articles related to those of an unacceptable risk. The criticism of the approach has been founded upon the focus of concentration on high risk machines more than others, in that this dismisses the complex of technology itself, and the impact that systems defined outside of this category could cause if used in unintended ways.[1187] The EDPB and EDPS also share their concern that there needs to be more alignment to the prevention of risk to fundamental rights alongside the GDPR, to ensure cohesiveness of the two regulations.[1188]

Even before the Act was passed, the failures of the risk-based approach were brought to light with the impossibility of categorising generative AI into the four risk categories set out. The allocation of risk is based on the intended use of systems, raising concerns when it comes to generative AI. Generative systems are not built for a specific context or purpose, and are commended for their openness to be used on an unprecedented scale, allowing end-users or data subjects to decide the intended use of the system.[1189] Due to this, it is difficult to state where generative systems should be placed on the risk spectrum if the intended use is unknown until it is in the hands of those affected by it, usually in the form of the public themselves.

[1185] Luca Bertuzzi, 'Leading MEPs raise the curtain on draft AI rules' (Euractiv Blog, 11 April 2022, Updated 13th April 2022) <https://www.euractiv.com/section/digital/news/leading-meps-raise-the-curtain-on-draft-ai-rules/> accessed 2nd January 2023.
[1186] European Commission (n 117) 3.
[1187] European Data Protection Supervisor (n 114) 13.
[1188] European Data Protection Board (n 782) 2.
[1189] Natali Helberger and Nicholas Diakopoulos, 'ChatGPT and the AI Act' [2023] 12(1) *Internet Policy Review* 1.

An example of a generative AI system is ChatGPT, software only in recent history freely accessible to the public; the system can do a wealth of tasks based on the prompts given by the end-user, meaning that depending on the information given, the system could fluctuate in its risk level, posing the question of which rules apply. To solve this issue, Helberger and Diakopoulos argue that generative AI needs to be included as a separate section outside of the risk allocation, with separate rules to be subjected to, or a combination of those already stated.[1190] This view is argued in an open letter to be unnecessary by Microsoft and other companies developing the technology,[1191] perhaps a convenient view taken to avoid over-interference with such systems. Interestingly, the US has also reportedly sided with the tech companies on this point, stating that general-purpose AI should be excluded from the Act, and that the regulation should focus on regulating a narrower definition of AI.[1192]

In response to this, the AI Act had to adapt to account for these models, producing further rules for generative AI which should be complied with in addition to the general obligations that must be followed, depending on the risk category of the system. The EP in their suggested amendments to the AI Act categorised generative AI as foundation models,[1193] which typically are designed on huge datasets, with generative AI having the ability to generate content from these datasets. The obligations for providers of general-purpose AI models are listed within Article 52(c) of the final draft of the AI Act, and include retaining up-to-date technical documentation and training records (like general high risk systems), putting in place respect for copyright policy, and publishing a detailed summary on the training data.[1194]

[1190] ibid.

[1191] Confederation of Industry of the Czech Republic, 'Open letter on the proposed regulation of artificial intelligence' (7th November 2022) <https://www.spcr.cz/images/Open_letter_on_the_proposed_regulation_of_artificial_intelligence_FIN20221107_125114.pdf> accessed 4th January 2023.

[1192] Luca Bertuzzi, 'The US unofficial position on upcoming EU Artificial Intelligence Rules' (Euractiv Blog, 24th October 2022, updated 26th October 2022) <https://www.euractiv.com/section/digital/news/the-us-unofficial-position-on-upcoming-eu-artificial-intelligence-rules/> accessed 8th January 2023.

[1193] European Parliament, *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))* (June 2023) Recital 60e.

[1194] Provisional Agreement for the AI Act (n 103) Article 52c.

In the final draft of the AI Act, the necessity to clarify the classification of general-purpose AI is noted in Recital 60(o), with an indication that general-purpose AI with systemic risks must adhere to stronger protocols.[1195] These additional rules are listed within Article 52(d), and include performing model evaluations, mitigating systemic risks and ensuring an adequate level of cybersecurity protection.[1196] Within the Recitals, it is stated that if a model meets the threshold for high-impact capabilities, then it would be classified as a general purpose AI model with systemic risks,[1197] which could include negative effects on critical sectors or infrastructures, to democratic processes, or the dissemination of illegal, false or discriminatory content.[1198] If a model fits into this category, in addition to complying with the obligations already included for AI and general-purpose systems, there are further rules, such as heightened testing through internal or independent means, and continuous efforts to mitigate the systemic risks.[1199] The late inclusion of these extra rules for general-purpose systems reveals the issues that stem from the risk-based approach to regulation.

The EC has set out to establish the European AI Office, one of the many authorities introduced to help enforce the rules and foster collaboration between institutions and agencies.[1200] The AI Office’s main responsibility is to oversee the advancements in AI models, particularly regarding generative AI, ensuring that the rules and obligations are being followed by the developers of such systems.[1201] The AI Office also has the task of fostering actions and policies in reference to prohibited practices, and to support the development of trustworthy AI systems.[1202] The Office intends to work collaboratively with stakeholders, other EU institutions, and bodies of MS on behalf of the EC.[1203] This Office can be seen as a positive step to help increase enforcement of the Act once in place, yet shows the additional support needed for systems that do not comfortably fit into the risk categorisation system.

[1195] ibid Recital 60o.
[1196] ibid Article 52d.
[1197] ibid Article 52d.
[1198] ibid Recital 60m.
[1199] ibid Recital 60q.
[1200] European Commission, *Commission Decision of 24.1.2024 establishing the European Artificial Intelligence Office* (390 final, 2024).
[1201] ibid 2.
[1202] ibid Article 2(2).
[1203] ibid Article 2(3), Article 4 and Article 5.

Looking at the categorisation system generally, the focus on the intended use of systems has drawn some apprehension for systems that could later be re-purposed for other uses. To give an example, the EP have focused their concerns on policing systems being repurposed to commit mass surveillance, and that such systems would fail on being proportionate and necessary to align with human rights protection,[1204] questioning whether there would be sufficient safeguards to protect from this once software has been placed on the open market. This concern has been echoed by several human right pressure groups, with the Future of Life Institute sharing their concerns on the consequences of the hidden intention of the uses of systems,[1205] and whether the scope of the regulation should also make considerations of the wider effects of such use. They emphasise that there must be explicit consideration of the impact on society at large, and give the example of ‘simple’ marketing applications being used to influence voting behaviour, which could then lead to effecting and manipulating election results, having a much bigger impact on society in comparison to the individual person.[1206]

These concerns were noted by the EP and were somewhat addressed, to account for scenarios in which systems are used for many different purposes, including where generative systems are integrated with other systems and adapt the risk level.[1207] The addition of rules for generative AI is welcomed, however due to the EU and the margin of appreciation, there is concern relating to how such rules for governing systems would be implemented into MS. It could be questioned that giving such discretion would lead to inconsistencies in application between different EU MS, making it technically difficult for such systems to operate according to different rules, particularly when systems are used online in the global sphere, like ChatGPT, a decision assumingly unwelcomed by tech companies.

Another area of criticism towards the Act was due to its broad exemptions given; that the obligations do not need to be adhered to for reasons of national security, and

[1204] European Parliament, ‘Artificial Intelligence in policing: safeguards needed against mass surveillance’ (European Parliament Press Release, June 2021) <https://www.europarl.europa.eu/news/en/press-room/20210624IPR06917/artificial-intelligence-in-policing-safeguards-needed-against-mass-surveillance> accessed 10th January 2023.

[1205] Future of Life Institute, *FLI Position Paper on the EU AI Act* (Feedback Reference: F2665546, August 2021) 4.

[1206] ibid 4.

[1207] Provisional Agreement for the AI Act (n 103) Recital 60d.

defence purposes.[1208] The national security exemption is a concept seen before in legislation, and is particularly concerning given that this gives scope for public bodies to avoid adhering to the statutory rules. The use of AI for military and defence purposes has also caused worry regarding the unethical use of systems.[1209] These vague exceptions provide substantial loopholes within the Act, and the possible abuse of these exceptions has been the basis to criticism of where the Act falls short.[1210] Due to this, there have been arguments that blanket exceptions should not exist, and that such systems must still undertake risk and impact assessments throughout their use, to uphold the rule of law.[1211]

The original proposal was also criticised due to its lack of focus on fundamental rights examination.[1212] This criticism was evidently noted, as the final draft of the AI Act introduced the obligations in Article 29a.[1213] Article 29a requires a FRIA before a high risk system is deployed, however, exceptions are provided, and the Article seems to focus only on systems that provide public services.[1214] Although FRIAs have long been sought after, the EU have missed the mark with this obligation. FRIAs should be mandatory for at least, all high risk systems, to ensure a human-centric approach is achieved. The obligation to conduct a FRIA only applies on the first use of the system, and includes a caveat that allows deployers to rely on previous FRIAs carried out by providers,[1215] which runs the risk of deployers not giving assessments the rigorous examination needed, and relying on previous assessments for the sake of reduced admin. Due to this, it is argued that this Article

[1208] ibid Article 2(3).
[1209] Arthur Holland Michel, 'Inside the messy ethics of making war with machines' (MIT Technology Review Blog, 16th August 2023) <https://www.technologyreview.com/2023/08/16/1077386/war-machines/> accessed 1st September 2023; Abhishek Gupta, 'Introduction to ethics in the use of AI in war' (Towards Data Science Blog, 24th February 2021) <https://towardsdatascience.com/introduction-to-ethics-in-the-use-of-ai-in-war-9e9bf8ba71ba> accessed 1st September 2023; European Parliament, *The ethics of artificial intelligence: Issues and initiatives* (European Parliamentary Research Service, March 2020) 63-65.
[1210] Francesca Fanucci and Catherine Connolly, 'What are the AI Act and the Council of Europe Convention' (Stop Killer Robots, 14th August 2023) <https://www.stopkillerrobots.org/news/what-are-the-ai-act-and-the-council-of-europe-convention/> accessed 1st September 2023; Civil Liberties Union for Europe and Others, 'Open letter: The AI Act Must Protect the Rule of Law' (Open Letter, September 2023) <https://dq4n3btxmr8c9.cloudfront.net/files/iytbh9/AI_and_RoL_Open_Letter_final_27092023.pdf> accessed 1st September 2023; Emre Kazim, Osman Güçlütürk, Denise Almeida, Charles Kerrigan, Elizabeth Lomas, Adriano Koshiyama, Airlie Hilliard and Markus Trengove, 'Proposed EU AI Act- Presidency compromise text: select overview and comment on the changes to the proposed regulation' [2023] 3 *AI and Ethics* 381.
[1211] Fanucci and Connolly (n 1215); Civil Liberties Union for Europe and Others (n 1215).
[1212] European Commission (n 117).
[1213] Provisional Agreement for the AI Act (n 103) Article 29a.
[1214] ibid Article 29a.
[1215] ibid Article 29a(2).

is inadequate, and needs further thought to put fundamental rights consideration at the forefront.

According to the proposal, the prohibited systems that were originally identified as an unacceptable risk were listed within Article 5.[1216] Article 5 listed the following systems; systems that materially distort a person's behaviour which is likely to cause physical or psychological harm,[1217] systems that exploit vulnerabilities of a specific group due to age, physical or mental disability, in order to materially distort behaviour and is likely to cause physical or psychological harm,[1218] social scoring systems used or placed on the market by public bodies in circumstances where the social score leads to detrimental or unfavourable treatment,[1219] and the use of 'real-time' remote biometric identification systems in public spaces by law enforcement, subject to some exceptions.[1220] The list comprises what the EU describe as systems that conflict with Union values, and clearly violate fundamental rights,[1221] which predominantly include the right to privacy, freedom of expression, and freedom from discrimination.[1222]

This risk category is deemed as necessary to ensure protection of fundamental rights and to respect the principles of proportionality and necessity, but questions are raised by many who state that the scope of these systems is too narrow, and that the exceptions listed within the Article will allow for clear abuse of such systems, whilst other terms may need further clarification. Article 5(1)(a) in particular needed to be clarified, which prohibited "*subliminal techniques beyond a person's consciousness to materially distort a person's behaviour*",[1223] leading to a likelihood of causing physical or psychological harm.[1224] Tech UK raise concern that it is unclear which systems this would extend to,[1225] and clarification is clearly needed.[1226] In the final

[1216] European Commission (n 117) Article 5.
[1217] ibid Article 5(1)(a).
[1218] ibid Article 5(1)(b).
[1219] ibid Article 5(1)(c).
[1220] ibid Article 5(1)(d).
[1221] ibid 12.
[1222] ECHR, Article 8, 10 and 14
[1223] European Commission (n 117) Article (5)(1)(a).
[1224] ibid Article (5)(1)(a).
[1225] TechUK (n 1185) 3.
[1226] ibid; Patrick Grady, 'EU's AI Act Resurrects Subliminal Messaging Panic' (Center for Data Innovation Blog, 21st October 2022) <https://datainnovation.org/2022/10/eus-ai-act-resurrects-subliminal-messaging-panic/> accessed 2nd September 2023.

draft of the Act, this has been adapted to the also include purposefully manipulative or deceptive techniques, which impairs a person's ability to make an informed decision, making them take a decision that they otherwise would not have taken, which in turn, causes significant harm.[1227] This adaption gives more context, but the concerns raised by Tech UK have arguably not been addressed;[1228] it remains unclear where this prohibition would extend to.

Social scoring systems are widely accepted to be a dystopian feature of society, with examples usually taken from China, who have a widespread social credit system in place that scores citizens based on certain actions, with a low score restricting them from certain luxuries, with some aspect of government involvement.[1229] However, similar systems in principle are widespread and accepted domestically; in the UK for example, several companies give credit scores to individuals, basing scores on previous actions by that individual, by past abilities to pay direct debits, which could lead to the rejection of opportunities. Social scoring is also widespread on several apps, including Uber, where individuals and drivers 'rate' each other based on their interactions, giving each individual an overall score based on their interactions.

The first proposal of Article 5 focused on social scoring by public authorities based on the trustworthiness of an individual based on their social behaviour or known or predicted personal characteristics, with such a score leading to detrimental or unfavourable treatment in social contexts, and/or treatment that is unjustified or disproportionate to their social behaviour or its gravity.[1230] This differentiation between public and private use meant that public social scoring systems were prohibited, whereas private use of systems would be considered high risk. Edwards has commented on the lack of clarity with public sector use, stating that it is not clear, with the example of whether assessing families for child abuse or neglect would fall under the scope of the Article.[1231] Interestingly, the final draft of the AI Act

[1227] Provisional Agreement for the AI Act (n 103) Article 5(1)(a).
[1228] TechUK (n 1185) 3.
[1229] Paul F Langer, 'Lessons from China – The Formation of a Social Credit System: Profiling, Reputation Scoring, Social Engineering' (The 21st Annual International Conference on Digital Government Research, pages 164-174, June 2020) <https://dl.acm.org/doi/10.1145/3396956.3396962> accessed 5th September 2023.

[1230] European Commission (n 117) Article 5(1)(c).
[1231] Lilian Edwards, *The EU AI Act: a summary of its significance and scope* (Ada Lovelace Institute, April 2022) 10.

has removed the differentiation between public and private, meaning that both public and private use of social scoring systems will be prohibited.[1232] This now raises interesting questions related to the future of credit checks and the other examples listed above, particularly when there is autonomy present in the decision-making process.

The majority of criticism towards this Article relates to Article 5(d), which prohibits "*the use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement*".[1233] EDRi argues that in reality, this only addresses a small range of practices, and does not prevent the use of systems for mass surveillance.[1234] Due to only 'real-time' systems being included, law enforcement can use systems retroactively to avoid this ban.[1235] The proposal of the AI Act listed a number of exceptions to the ban under Article 5,[1236] including when such systems are necessary for the points listed below. These exceptions were referred to as major loopholes, allowing the risk of abuse to take place from law enforcement due to the broadness and vagueness of the exceptions. However, the final draft of the AI Act has made some welcomed adaptions to these. In addition, to address these concerns, the final draft of the AI Act includes a provision for a FRIA before 'real-time' remote biometric identification systems are lawful to be used in publicly accessible spaces,[1237] which is supported.

- "*(i) the targeted search for specific potential victims of crime, including missing children;"*[1238]

This first exception has been adapted to the targeted search for victims of abduction, trafficking, sexual exploitation, and missing persons.[1239] This exhaustive list has narrowed the scope of the exception, seemingly addressing the worry of the possible abuse that the proposed exception would have allowed.

[1232] Provisional Agreement for the AI Act (n 103) Article 5(1)(c).
[1233] ibid Article 5(1)(d).
[1234] EDRi, 'Remote Biometric Identification: a technical and legal guide' (EDRi website, 23rd January 2023) <https://edri.org/our-work/remote-biometric-identification-a-technical-legal-guide/> accessed 11th September 2023.
[1235] Sebastian Klovig Skelton, 'Europe's proposed AI regulation falls short on protecting rights' (Computer Weekly Blog, 14th June 2021) <https://www.computerweekly.com/feature/Europes-proposed-AI-regulation-falls-short-on-protecting-rights> accessed 12th September 2023.
[1236] European Commission (n 117) Article 5(1)(d)(i-iii).
[1237] Provisional Agreement for the AI Act (n 103) Article 5(2).
[1238] European Commission (n 117) Article 5(d)(i).
[1239] Provisional Agreement for the AI Act (n 103) Article 5(1)d(i).

- *"(ii) the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or of a terrorist attack;"*[1240]

The wording of this second exception in the final draft of the AI Act has also taken into consideration the concerns towards possible abuse and broadness of scope, giving the additional context of "*a genuine and present or genuine and foreseeable threat of a terrorist attack*".[1241]

- *(iii) the detection, localisation, identification or prosecution of a perpetrator or suspect of a criminal offence referred to in Article 2(2) of Council Framework Decision 2002/584/JHA 62 and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least three years, as determined by the law of that Member State.*"[1242]

The third exception has been reworked in the final draft of the AI Act as follows, most notably removing the reference to the external Council decision, and instead, including a list of offences, in the newly produced Annex IIa: "*the localisation or identification of a person suspected of having committed a criminal offence, for the purposes of conducting a criminal investigation, prosecution or executing a criminal penalty for offences, referred to in Annex IIa and punishable in the MS… for a maximum of at least four years.*"[1243] It should also be noted the increase from three to four years in terms of the offences included, increasing the range and applicability of the system.

Regardless of these changes, there have been mass calls to ban biometric identification in public spaces outright,[1244] to ensure adequate protection of privacy,[1245] particularly after several reports of the extent to which systems have been used disproportionately around the EU.[1246] The EDPB and EDPS support the argument for a blanket ban of automated identification in publicly accessible spaces regardless of the context, stating that the wording used within Article 5 is flawed, and would only apply in exceptional cases.[1247] The EDPB and EDPS also argue that AI

[1240] European Commission (n 117) Article 5(d)(ii).
[1241] Provisional Agreement for the AI Act (n 103) Article 5(1)d(ii).
[1242] European Commission (n 117) Article 5(d)(i-iii).
[1243] Provisional Agreement for the AI Act (n 103) Article 5(1)d(iii).
[1244] Reclaim Your Face, 'Reclaim Your Face' (Website) <https://reclaimyourface.eu/> accessed 13th September 2023.
[1245] ECHR (n 3) Article 8.
[1246] EDRi, The Rise and Rise of Biometric Mass Surveillance in the EU (2021).
[1247] European Data Protection Board (n 782) 10-12.

systems relying on biometrics to categorise individuals into groups based on ethnicity, gender, political or sexual orientation should be outright banned, to ensure full protection against discrimination and against uses of AI that are not scientifically feasible,[1248] like identifying political orientation based on biometrics; biometrics of an individual remain the same where political opinions can change.

Predictive policing took particular focus during the debates on the proposal, where the EP had previously noted these systems as undermining human dignity and the presumption of innocence, in addition to the risk of discrimination, and hence, argued for amendment of the AI Act for such systems to fall within Article 5's unacceptable risk systems and be banned.[1249] In addition to this, EDRi also calls for a ban of AI used by law enforcement or criminal justice to predict and profile behaviour, and those used to control migration, stating that these uses are incompatible with fundamental rights and the notion of democracy.[1250]

Several MEPs also raise the issue of mass surveillance in relation to predicting policing and police use of AI in general, posing an infringement to Article 8 and Article 14,[1251] where identification systems have several reports of reflecting inaccuracies, particularly to minority ethnic groups, those who identify as LGBTQ+, among other groups.[1252] The use of such systems, particularly by law enforcement and the judiciary has the risk of amplifying existing discrimination, and in turn, would not promote public trust in the use of AI, and if regulated in a similar way in the UK, would pose domestic concerns of breaches to the Equality Act,[1253] reversing the intentions of the AI Act itself. It should be recognised that not all applications are necessary and proportionate for a democratic society, particularly a society that is committed to promoting and upholding human right protections.[1254]

[1248] ibid 12.

[1249] European Parliament, *Draft Report on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM2021/0206 – C9-0146/2021 – 2021/0106(COD))* (2022) 54.

[1250] EDRi (and 44 others) (n 739).

[1251] ECHR (n 3) Article 8 and 14.

[1252] European Parliament (n 1209); Catriona Campbell, 'The EU's Draft AI Regulations: A Thoughtful But Imperfect Start' (Medium Blog, 13th August 2021) <https://catriona-campbell.medium.com/the-eus-draft-ai-regulations-a-thoughtful-but-imperfect-start-8c6489f1617> accessed 15th September 2023.

[1253] Equality Act 2010 s4.

[1254] Skelton (n 1240).

These concerns have clearly been taken into consideration by the EU institutions during the debate process of the AI Act, and the list included within Article 5 has been expanded upon, a move by which this thesis supports. In reference to the above discussion, Article 5 in the final AI Act now also includes "*use of an AI system for making risk assessments… to assess or predict the risk… of a natural person to commit a criminal offence, based solely on the profiling of a natural person or on assessing their personality traits and characteristics.*"[1255] In principle, this Article could be seen as touching upon the area of predictive policing, however it also includes the caveat that this prohibition will not apply to systems used to support human assessment, which is said to be "*already based on objective and verifiable facts directly linked to a criminal activity"*[1256] and hence, leaves a loophole for predictive policing systems to be used. Due to this, the concerns towards predictive policing still stand, particularly when systems that are not included in this ban have been shown to reinforce existing police discrimination.[1257]

Previously, the EP's Civil Liberties Committee called for an extension of the ban, to cover law enforcement using private facial recognition databases,[1258] such as Clearview AI, which was the basis of a mass data protection scandal and fined accordingly by the ICO in 2022.[1259] In the final draft of the AI Act, this call has been addressed.[1260] This comes in the form of Article 5 now also including AI systems that are placed onto the market for the specific purpose of "*creating or expanding facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage*".[1261] This inclusion is welcomed, and addresses the concerns raised through the Clearview scandal.[1262]

The final draft also includes emotion recognition systems in the list of prohibited practices, when used in workplace and educational institutions, except for use for

[1255] Provisional Agreement for the AI Act (n 103) Article 5(1)(da).
[1256] ibid Article 5(1)(da).
[1257] Fair Trials, *Automating Injustice* (2021) 27-30.
[1258] European Parliament (n 1209).
[1259] ICO, 'ICO fines facial recognition database company Clearview AI Inc more than £7.5m and orders UK data to be deleted' (Press Release, 23rd May 2022) < https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/05/ico-fines-facial-recognition-database-company-clearview-ai-inc/> accessed 12th September 2023.
[1260] European Parliament (n 1209).
[1261] Provisional Agreement for the AI Act (n 103) Article 5(1)(db).
[1262] ICO (n 1264).

medical or safety reasons.[1263] Again, the exception to these systems is unnecessarily broad, and may need further interpretation to ensure the prohibition works as intended. In principle, the expansion of the list shows a step forward in the approach to recognise the dangers AI systems can present, however as shown, there remain several loopholes and unclear scopes, which could have been improved upon further.

The systems defined as 'high risk' take the proposal's focus, and are subject to several mandatory requirements within the legislation. The classification of a high risk system is based on the intended purpose of the machine, rather than the function it performs, with two cumulative criteria given in Article 6, quoted above and paraphrased here; as systems intended to be used as a safety component or products themselves covered by legislation listed within Annex II of the legislation,[1264] and required to conform to third-party conformity assessments,[1265] in addition to systems that are explicitly listed in Annex III.[1266]

To give some examples of systems classified as high risk, Annex III includes biometric and categorisation systems intended to be used for identification of humans,[1267] systems intended to give access to education or training,[1268] and systems used by law enforcement for predicting criminal offences based on profiling,[1269] in addition to several others. This list contains systems where the risks are clear, or that pose risks which are likely to materialise in the near future, which shows the EU's forward-thinking approach. Article 7 of the Act states that the EC should amend the list in line with Article 73[1270] in circumstances where systems are intended to be used in any of the already listed areas, or if a system poses a risk to health and safety, or an adverse impact on fundamental rights equivalent or greater than those systems already listed.[1271] For the AI Act to achieve its intended human-centric status, the EP argue that some high risk systems have not been recognised

[1263] Provisional Agreement for the AI Act (n 103) Article 5(1)(dc).
[1264] ibid Article 6(1)(a).
[1265] ibid Article 6(1)(b).
[1266] ibid Article 6(2).
[1267] ibid Annex III, 1(a).
[1268] ibid Annex III, 3(a).
[1269] ibid Annex III, 6(e).
[1270] ibid Article 7 and 73.
[1271] ibid Article 7.

within the Act for their real risks to fundamental rights, and need to fall outside of the high risk category and into the unacceptable category.[1272]

The standards to determine whether a system is high risk are overly broad, and the wording of the Article could encompass systems that are not intended to fall into the category, causing unnecessary restrictions and costs for those who develop such systems.[1273] There have been attempts to address this concern within the final draft of the AI Act, where Article 6(2)(a) has a provision that allows derogation from the high risk category if it can be shown that a significant risk of harm does not exist.[1274] This intends to be measured by fulfilling one or more criteria, set out within the Act, and includes where a system is intended to perform a narrow procedural task,[1275] or where the system intends to improve the result of an activity previously completed by humans,[1276] in addition to several other factors.

Title III, Chapter Two of the AI Act, from Article 8 to Article 15, details the requirements high risk systems need to comply with, including rules on risk management, data governance, documentation, record keeping, transparency, human oversight, robustness, accuracy and cybersecurity. [1277] The EU describe the rules as proportionate and necessary to achieve the objectives of the regulation, and that a sufficient regulatory burden to systems that pose a high risk to fundamental rights and safety is imposed.[1278] Article 8 of the AI Act places an obligation of compliance with the chapter for high risk systems, and sets out that compliance will be measured based on the intended purpose of the system, in addition to the risk management system addressed in Article 9.[1279] The risk management system should run continuously throughout the lifecycle of a high risk system and be documented, comprising of the following:

- Identification and analysis of associated known and foreseeable risks;[1280]

[1272] European Parliament (n 1254) 54.
[1273] TechUK (n 1185) 1.
[1274] Provisional Agreement for the AI Act (n 103) Article 6(2a).
[1275] ibid Article 6(2a)(a).
[1276] ibid Article 6(2a)(b).
[1277] ibid Article 8-15.
[1278] European Commission (n 117) 7.
[1279] Provisional Agreement for the AI Act (n 103) Article 8 and Article 9.
[1280] ibid Article 9(2)(a).

- Estimation and evaluation of such risks when the system is used in line with its intended purpose, and under reasonably foreseeable misuse;[1281]
- Evaluation of other possible risks based on analysis of data gathered from post-market monitoring in line with Article 61;[1282]
- And adoption of suitable risk management measures to address the risks; if risks cannot be eliminated, adequate mitigation plans should be put in place.[1283]

The compliance with Article 9 from a business perspective adds an extra layer of complexity as not only identifiable risks need to be mitigated, but also reasonably foreseeable ones.[1284] There is also reference to foreseeable misuse, defined under Article 3(13) within the AI Act as "*the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other systems*".[1285] Whether foreseeability will be judged from a subjective or objective test will be a decision for the Courts, as it is likely that this will come under scrutiny in the future, but it would be assumed that an objective basis would be the fairest.

Article 10 of the AI Act governs data and data governance; applying to high risk systems that are trained through datasets.[1286] The training, validation and testing of datasets should meet the criteria within the Article, which include ensuring that datasets are relevant, representative, free of errors and complete.[1287] It is not entirely clear how simple compliance with this would be, particularly in terms of representation and ensuring datasets are free of errors, and further clarification or techniques to assess these areas may be needed to ensure consistency within machines. For datasets to be representative and for bias to be identified, more data is needed to be able to detect patterns, and there may be issues with the data protection principle of data minimisation under the GDPR, Article 5,[1288] which would

[1281] ibid Article 9(2)(b).
[1282] ibid Article 9(2)(c).
[1283] ibid Article 9(2)(d).
[1284] Mark Cankett and Barry Liddy, 'Risk management in the new era of AI regulation' (Deloitte Blog: Audit & Assurance, 12th July 2022) <https://www2.deloitte.com/uk/en/blog/auditandassurance/2022/the-new-era-of-ai-regulation.html> accessed 14th September 2023.
[1285] Provisional Agreement for the AI Act (n 103) Article 3(13).
[1286] ibid Article 10.
[1287] ibid Article 10(3).
[1288] GDPR (n 1) Article 5.

need further clarification for how the two regulations would align. Conflict with the GDPR can also be seen within Article 10(5) of the AI Act,[1289] whereby it is stated that special categories of data can be processed for bias monitoring, detection, and correction within high risk systems,[1290] meaning adaption of the exceptions to process special category data under the GDPR need to be made to align with this under Article 9(2).[1291]

Article 11 of the AI Act governs technical documentation, ensuring such documentation is drawn up before a system is on the market or in service, and that documentation is kept up to date.[1292] This technical documentation should contain all the necessary information to assess and evidence compliance, and should be used as proof for enforcement bodies if businesses are notified.[1293] Article 12 aligns with this, ensuring systems have record-keeping capabilities, to allow traceability and assessment of the functioning of machines through their lifecycle, in which the records can be used to ensure compliance with monitoring requirements.[1294] Article 13 governs the rules on transparency, with an obligation for high risk systems to be designed and developed to ensure their output can be interpreted and used appropriately, and that instructions for use should be made to ensure systems are accessible and comprehensible for users.[1295] Article 14 governs mechanisms of human oversight, to prevent and minimise risks to health, safety and/or fundamental rights,[1296] and Article 15 oblige systems to be developed to achieve appropriate levels of accuracy, robustness and cybersecurity.[1297]

The AI Act provides obligations on providers and deployers of AI systems, to ensure compliance with the requirements, the most important to note within Article 43 governing conformity assessments.[1298] Providers must ensure systems undergo a conformity assessment prior to systems being on the market or in service,

[1289] Artur Bogucki, Alex Engler, Clément Perarnaud and Andrea Renda, *The AI Act and Emerging EU Digital Acquis: Overlaps, gaps and inconsistencies* (CEPS, September 2022) 6-7.
[1290] Provisional Agreement for the AI Act (n 103) Article 10(5).
[1291] GDPR (n 1) Article 9(2).
[1292] Provisional Agreement for the AI Act (n 103) Article 11(1).
[1293] ibid Article 11(1).
[1294] ibid Article 12.
[1295] ibid Article 13.
[1296] ibid Article 14.
[1297] ibid Article 15.
[1298] ibid Article 43.

demonstrating compliance with Chapter Two, in addition to adhering to an EU declaration of conformity under Article 48, and CE marking governed under Article 49.[1299] To make completion of conformity assessments more accessible, tools have been developed to give practical guidance to those fulfilling assessments.[1300]

Under Article 48, the provider must draw up an EU declaration of conformity for the system in writing once it is completed, and retain it for 10 years to be used by national authorities if requested.[1301] Under Article 49, systems that have satisfied the conformity assessment requirements must adhere to a mandatory CE-marking procedure,[1302] and be registered on an EU-wide database,[1303] which then allows such system to be placed on the market or be put in service in European markets. The EDPB and EDPS comment that further alignment with the CE-marking procedure and data protection certificates should be made, allowing collaboration between authorities of both regulations and harmonising standards.[1304]

Once on the market or in service, providers of high risk AI systems need to comply with Article 61, which governs post-market monitoring rules.[1305] To oblige with this Article, providers must establish and document a post-market monitoring system proportionate to the risk the system poses, with the ability to collect, document and analyse relevant data through the lifecycle of the system, to ensure continuous compliance with the rules in Title III, Chapter Two.[1306] The AI Act states that the EC intends to adopt an implementing Act to guide providers, containing a template for a post-market monitoring plan with the elements to be included, which needs to be set out as part of the technical documentation prior to the system being on the market, as part of complying with Article 11,[1307] and put into practice post-market, to comply with Article 61.[1308]

[1299] ibid Article 48 and Article 49.
[1300] Floridi, Holweg, Taddeo, Silva, Mökander and Wen (n 1189).
[1301] Provisional Agreement for the AI Act (n 103) Article 48.
[1302] ibid Article 49.
[1303] ibid Article 51 and Article 60.
[1304] European Data Protection Board (n 782) 3 and 9-10.
[1305] Provisional Agreement for the AI Act (n 103) Article 61.
[1306] ibid Article 61.
[1307] ibid Article 11.
[1308] ibid Article 61.

Other systems that do not fall into the unacceptable or high risk category are defined as low risk, and are subjected only to minimal transparency obligations depending on their purpose and intended use.[1309] The low risk category is divided into two; those which pose a limited risk, and those that pose minimal risk. Limited risk systems must comply with obligations set out under Article 52 related to transparency,[1310] whereby systems classified as minimal risk can voluntarily comply with impact assessments and codes of conduct.[1311] Limited risk systems are those that interact with humans, in which humans should be informed they are interacting with AI, unless it is deemed obvious from the circumstances and context of the use,[1312] which is likely to lead to global disclosure on most websites and apps.[1313]

Access Now has raised concern about the low risk categorisation, arguing that the proposal fails to adequately protect fundamental rights, particularly concerning biometric applications such as emotion recognition and AI polygraphs, and suggests that the Act has categorised these incorrectly, and that they should fall under the unacceptable risk category.[1314] Using the example of deepfakes; the Act itself categorises these as limited risk, but arguably, the use of deepfakes could, depending on context, fall within Article 5(1)(a) and (b),[1315] if capable of manipulating and distorting those who are subject to the deepfake, and therefore be banned, however this is unclear within the Act. The categorisation and sliding scale applicable to the context of use are not captured within the risk categorisation system, which reveals another example of the issues in the fundamental approach to the regulation.

The Act will be enforced through a governance mechanism imposed through MS, allowing states to build on structures that already exist, and ensure cooperation through the introduction of a Union-level organisation in the form of a 'European Artificial Intelligence Board' (EAIB).[1316] In addition to this, the 'European AI Office'

[1309] ibid Article 52.
[1310] ibid Article 52.
[1311] ibid Article 69.
[1312] ibid Article 52.
[1313] Engler (n 1148).
[1314] Access Now, *Access Now's submission to the European Commission's adoption consultation on the Artificial Intelligence Act* (Feedback Reference: F2665462, August 2021) 9-10.
[1315] Provisional Agreement for the AI Act (n 103) Article 5(1)(a) and (b).
[1316] ibid Article 56.

will be established, with a focus on enforcing the rules on general-purpose AI models and cooperating with the Board on AI policy and collaboration between institutions.[1317] Within MS, enforcement bodies will be established, similar to the oversight mechanism within the GDPR. Following Brexit, it will be interesting to see whether an independent body is established for AI in the UK like the ICO, or that the UK will attempt to set itself apart from EU rules by aligning roles within pre-existing bodies.

The EDPB and EDPS argue that the EAIB should be given more autonomy to ensure consistency of application of the regulation within the MS, as currently, the Act suggests that the EC will be involved with the Board, denying the need for an authority body that is free from political influence, and in turn, the clarification needed regarding the legal status of the EAIB.[1318] The University of Cambridge commented that an adaption to the AI Act to allow the EAIB to add to the Annexes would assist in ensuring the regulation keeps up-to-date with the advancements in AI, and that changes can be made on the review of enforcement issues noted by the Board.[1319] Under Article 71, the penalties for breaching the proposal are clear (for breaches of Article 5, €35 million or 7% of annual turnover, whichever is higher, otherwise €15 million or 3% of annual turnover, whichever is higher),[1320] and follow a similar approach to the GDPR penalties,[1321] however, there is a lack of discussion on redress mechanisms and how individuals can claim under the AI Act, differing substantially from the GDPR.[1322]

On reflection of the detailed examination of the AI Act, it is also important to note whether it is fit for purpose for real-life scenarios, and hence the next section will apply the Act in its current form to an example case study to reflect on its capabilities.

[1317] European Commission (n 1205).
[1318] European Data Protection Board (n 782) 15-16.
[1319] University of Cambridge, Submission of Feedback to the European Commission’s Proposal for a Regulation laying down harmonised rules on artificial intelligence (Feedback Reference: F2665626, August 2021) 6.
[1320] Provisional Agreement for the AI Act (n 103) Article 71.
[1321] GDPR (n 1) Article 83.
[1322] Skelton (n 1240).

5.2.2 AI Act Case Study

This section of the chapter will assess the proposals within the AI Act from the perspective of a case study, evaluating the rules relating to both unacceptable and high risk systems, using the foundation of the facts from the *R(Bridges)* case, focused on the use of FRT by law enforcement. The use of case studies seeks to demonstrate the practical application and impact of the regulatory responses made by the EU. Through examining real-world examples, this section aims to assess how the current EU approach would address the challenges highlighted in these cases, giving an indication to their effectiveness. These case studies are chosen not only due to highlighting and representing critical moments of public and legal scrutiny, but also because they offer a lens to evaluate the EU approach. Through comparing the facts of the *R(Bridges)* case with the rules under the AI Act, this section will illustrate how the new framework could provide more robust protection for fundamental rights, highlighting the need and benefits of a new regulatory framework, which the thesis advocates for.

The R(Bridges) FRT Scenario

The FRT involved in the *R(Bridges)* case was a pilot project known as 'AFR Locate', which involved the capturing of digital images of members of the public, which were processed and compared by the technology with digital images that made up the then current 'watchlist' of South Wales Police.[1323] 'AFR Locate' is a real-time remote automated facial recognition (AFR) system,[1324] used to assess live camera feeds in the moment and hence, falls into the definition under Article 3(37) of the AI Act proposal of a "*real-time remote biometric identification system*".[1325] In terms of classifying the system under the proposal, the use would fall under Article 5(d): "*the use of real-time remote biometric identification systems in publicly accessible spaces for the purposes of law enforcement*",[1326] being an unacceptable risk system, and hence banned. However, as noted in the above section, Article 5(d) contains several exceptions, including the targeted search for specific victims of crime, the prevention

[1323] *R(Bridges) v CC of South Wales Police* (n 47) [7].
[1324] Bethan Davies, Martin Innes and Andrew Dawson, *An Evaluation of South Wales Police's Use of Automated Facial Recognition* (Cardiff University, September 2018) 4.
[1325] Provisional Agreement for the AI Act (n 103) Article 3(37).
[1326] ibid Article 5(1)(d).

of specific threats to life, physical safety or terrorist attacks, or more notably for this case study, (and using the most recent phrasing of the exception) "*the localisation or identification of a person suspected of having committed a criminal offence… referred to in Annex IIa… and punishable in the MS concerned by a custodial sentence or a detention order for a maximum period of at least four years*."[1327]

South Wales Police contended that AFR was deployed to locate and detain wanted 'priority and prolific offenders', including a 'red' watchlist including perpetrators of serious crime, an 'amber' watchlist comprising of offenders wanted on warrant, and a 'purple' watchlist containing suspects of crimes in the area, ranging from summary to indictable offences.[1328] The offences listed within Annex IIa include mostly indictable offences, ranging from human trafficking to GBH,[1329] meaning that the exception of using the FRT would apply for the 'red', 'amber' and 'purple' watchlists, but depending on the crimes committed. In those cases where the crimes fall outside of the exception (and the crimes listed in Annex IIa), the system would be prohibited from use, meaning that it would be the responsibility of the deployer, in this case, South Wales Police, to ensure the data of those offences that align with the exception are filtered into the system and that the other offenders are not, which in turn, reflects the training and education needed for deployers of systems, to ensure they are used in the 'correct' circumstances.

For those circumstances in which the exception applies, the technology would need to adhere to the high risk requirements, meaning that before the system could be placed on the market in the EU-sphere, it would have to undergo the conformity assessment within Article 43,[1330] the EU declaration under Article 48,[1331] CE-marking procedure under Article 49,[1332] a FRIA under Article 29a,[1333] and be registered on the EU database.[1334] Based on the information given within the *R(Bridges)* case, and applying the AI Act, in terms of whether the system used by South Wales Police would satisfy the conformity assessment requirements, the system would have

[1327] ibid Article 5(1)(d)(iii).
[1328] *R(Bridges) v CC of South Wales Police* (n 47) [11] and [14].
[1329] Provisional Agreement for the AI Act (n 103) Annex IIa.
[1330] ibid Article 43.
[1331] ibid Article 48.
[1332] ibid Article 49.
[1333] ibid Article 29a.
[1334] ibid Article 60.

arguably breached Article 12, in addition to the post-market obligations under Article 61.[1335] The technology used by South Wales Police automatically and immediately deleted any person who did not, according to the system, match with a person of interest on one of the watchlists,[1336] and hence the record-keeping of the system is limited to 'correct' matches, reducing the ability to assess the functioning of machines through examining the likelihood of false negatives passing through the system, one of the notions used in measuring accuracy.

In short, the AI Act would allow FRT in the instance of *R(Bridges)* to be used in circumstances falling under the exception, contrasting the decision of the Court of Appeal in 2020 due to the prescribed by law requirement being satisfied by the AI Act. Depending on whether the system hypothetically conformed to the requirements for high risk systems, the user, in this case, South Wales Police, would perhaps be penalised, but the question remains of whether this is adequate in protecting rights. Essentially this case reflects the use of FRT to arguably align with the risks of AI for mass surveillance, and the concerns behind law enforcement use of systems, with the repercussions not equating to the risks posed.

5.2.3 The Proposed AI Liability Directive

In the EC's Communication on AI released back in 2018, in addition to the proposal for the established framework for AI, the EU raised concerns regarding the suitability of current rules in answering civil law liability questions involving the use of AI.[1337] This led to a subsequent EC report introduced in 2020 on the safety and liability implications of AI,[1338] which aimed to establish the potential gaps in liability frameworks for the new complexities brought by AI,[1339] which were discussed in detail in Chapter Three of this thesis, and include issues with foreseeability, the establishment of a clear duty of care and establishment of fault. The 2020 report highlights suggestions made in 2019 related to reversing the burden of proof in an adaption of national laws to provide a solution to AI-related damage,[1340] but at the

[1335] ibid Article 12 and 61.
[1336] *R(Bridges) v CC of South Wales Police* (n 47) [16].
[1337] European Commission (n 1144) 3.3.
[1338] European Commission (n 1165).
[1339] ibid 7.
[1340] ibid 14.

time it was questioned whether this was a sufficient permanent solution, or a short-term solution. The 2020 report concluded with emphasis on the new challenges brought to current liability regimes through the use of AI and that there is a pressing need to address these challenges to ensure there exists sufficient protection and compensation mechanisms for victims, whilst also maintaining technological innovation.[1341]

These conclusions have since been potentially addressed with the introduction of the proposed AI Liability Directive in 2022,[1342] which makes the EU's stance clear that current national liability rules, particularly those based on fault, are inadequate in dealing with AI-related claims,[1343] a stance also made clear by the EP.[1344] Due to this inadequacy, there was a clear need for updated rules that can be applied in a harmonised manner, to ensure legal certainty and encourage a consensus across the EU on civil liability, to promote the protection of victims from AI-related incidents. The objective of the proposed AI Liability Directive is to primarily promote trustworthy AI, through ensuring victims of damage can obtain equivalent protection to those in 'non-AI' scenarios, and to prevent a fragmented approach to civil liability across the EU.[1345] The proposed Directive forms part of the EU package to regulate AI, and runs alongside the AI Act to complement and achieve the overall aims of protecting fundamental rights.[1346]

It is important to note the scope of the proposed AI Liability Directive, in that it intends to be a targeted approach to national liability claims predominantly based on the fault of any individual, with the view of compensating any type of damage and any type of victim. This Directive is complemented by proposed adaptions to the PLD, which covers no-fault liability for defective products, which leads to damage suffered.[1347] The proposed changes include easing the burden of proof in complex

[1341] ibid 16-17.

[1342] Proposed AI Liability Directive (n 104).

[1343] ibid 1.

[1344] European Parliament, *Resolution of 3 May 2022 on artificial intelligence in a digital age (2020/2266(INI))* (2022).

[1345] Proposed AI Liability Directive (n 104) 2.

[1346] ibid 2.

[1347] European Parliament, 'New Product Liability Directive' (Briefing, September 2022) <https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739341/EPRS_BRI(2023)739341_EN.pdf> accessed 20th September 2023.

cases (which could include cases involving AI), and ensuring legal certainty to ensure better alignment with the digital era.[1348]

This approach to revise the liability rules is commented on by the EC as a holistic approach, to sufficiently address the issues related to AI to ensure adequate protection for future victims of AI-related claims.[1349] Public trust in AI has commonly been highlighted as one of the biggest ethical issues to address to improve future success of the technology, and clarification on liability rules is a step in the right direction to give reassurance to society that there are clear rules in place for redress where damage has been caused.

The proposed AI Liability Directive is based upon two drastic changes; the introduction of a rebuttable dual presumption, one relating to causation, and one on the existence of the causal link between fault and damage,[1350] and power given to national courts to order disclosure of evidence about high risk systems that are suspected of having caused damage.[1351] The Directive follows the definitions within the AI Act to ensure consistency between the two regulations, and intends to work on a complementary basis to the AI Act. Under Article 3 of the proposed Directive, domestic courts would have the power to order disclosure of evidence related to high risk systems, enabling victims to access information to identify who should be held liable.[1352] It is important to note that under the Directive, disclosure of evidence must be necessary and proportionate,[1353] and if such disclosure is not complied with by the associated provider or user of the system, it will be viewed as non-compliance with the duty of care obligations.[1354]

Article 4 of the Directive introduces a rebuttable presumption, giving claimants a more reasonable burden of proof in comparison to usual national rules, and a fairer chance at succeeding in liability claims. The rebuttable presumption does not apply

[1348] European Commission, *Proposal for a Directive of the European Parliament and of the Council on liability for defective products COM 2022/0302(COD)* (495 final, 2022) 2.
[1349] Proposed AI Liability Directive (n 104) 3.
[1350] ibid Article 4.
[1351] ibid Article 3(1).
[1352] ibid Article 3.
[1353] ibid Article 3(4).
[1354] ibid Article 3(5).

automatically; several conditions need to be met for the presumption to be triggered, which include:

(i) The claimant needs to show, or through presumption by the court, that the non-compliance with an EU or national obligation relevant to the harm amounts to a breach of a duty of care,[1355]

(ii) That it be reasonably likely, based on the circumstances, that the defendant's negligent conduct has influenced the output that gave rise to the damage and,[1356]

(iii) The claimant needs to show that the output produced by the system gave rise to the damage.[1357]

These conditions reduce the typical burden of proof for liability claims, but it should be made clear that this is not a reversal of the burden of proof, and that the burden still relies on the claimant to show the breach that has occurred. This Article has been criticised by the Ada Lovelace Institute, who suggest that a complete reversal of the burden of proof would be more adequate in clarifying the rules and scope of liability, which in turn would allow greater legal certainty.[1358] It should also be emphasised that the above approach only applies to systems categorised as high risk under the AI Act, and for those systems falling outside of the category, the presumption would only apply where domestic courts consider it excessively difficult for the claimant to prove the causal link.[1359]

5.2.4 AI Liability Case Study

This section of the chapter will examine the proposals within the AI Liability Directive through use of a case study, using the basis of the facts from the Government's use of the Office of Qualifications and Examinations Regulation's (Ofqual) A-Level algorithm to assess grades during the COVID-19 pandemic, which subsequently resulted in major media and public outrage, and eventually, a Government U-turn in

[1355] ibid Article 4(1)(a).
[1356] ibid Article 4(1)(b).
[1357] ibid Article 4(1)(c).
[1358] Christiane Wendehorst, *AI Liability in Europe: anticipating the EU AI Liability Directive* (Ada Lovelace Institute, September 2022) 16.
[1359] Proposed AI Liability Directive (n 104) Article 4(5).

2020.[1360] This real-world scenario reflects the need for a clear liability framework to allocate responsibility when AI systems fail to meet fairness and accuracy standards. By applying the facts of this scenario to the proposals within the AI Liability Directive, this section reflects how regulatory reforms could handle similar incidents. The case study demonstrates the practicality of the proposed framework and underscores the need for regulatory intervention to prevent harm and uphold human rights. This directly ties into the thesis' argument for a more proactive and robust regulatory framework that protects fundamental rights, whilst also ensuring developers and users of AI are responsible for their systems.

The Government's A-Level Grading Algorithm Scenario

Ofqual's 'A-Level Grading' algorithm was relatively simple, and did not involve ML or similar techniques, but instead was a statistical model used to calculate grades without human involvement. To re-emphasise the earlier discussion, the proposed AI Act originally defined 'AI systems' broadly as "*software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with*".[1361] Applying this definition, Ofqual's algorithm would fit within the scope of the proposed Act as it used a statistical approach, which was a technique listed within Annex I(c),[1362] and generated outputs of decisions in the form of grading students. Applying the 'new and improved' definition,[1363] as there was no human involvement with the decision-making of the algorithm, it can be argued that there is a level of autonomy, and an output is made based on inputs received, so the more recent definition would arguably still capture this algorithm.

The 'provider' of this system would be Ofqual, a non-ministerial government department with jurisdiction in England,[1364] as they developed the system under their name for use throughout the education sector. The relevant exam boards would be

[1360] Adam Harkens, 'Not just A-levels: unfair algorithms are being used to make all sorts of government decisions' (The Conversation, 3rd September 2020) <https://theconversation.com/not-just-a-levels-unfair-algorithms-are-being-used-to-make-all-sorts-of-government-decisions-145138> accessed 18th September 2023.
[1361] European Commission (n 117) Article 3(1).
[1362] ibid, Annex I(c).
[1363] Provisional Agreement for the AI Act (n 103) Article 3(1).
[1364] Office of Qualifications and Examinations Regulation (Ofqual), 'About us' (Government website) <https://www.gov.uk/government/organisations/ofqual/about> accessed 19th September 2023.

considered the 'deployer' of the system, as they used the system under their authority, and hence both Ofqual and the exam boards would be subject to the rules and obligations laid down within the AI Act. The algorithm used would be considered high risk under Article 6,[1365] falling under the specified list in Annex III, 3(b) on educational and vocational training, which includes "*AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to educational institutions*."[1366] Due to this, Ofqual would have to comply with the obligations under the Act for high risk systems,[1367] whilst the exam boards would be bound by Article 29,[1368] in that they should use the system in accordance to the instructions laid out by Ofqual.

The use of the algorithm during the COVID-19 pandemic back in 2020 quickly resulted in threatened legal action, backed by the Good Law Project, after receiving mass reports from students whose grades had been downgraded, resulting in the loss of university places, job offers, funding and scholarships.[1369] The legal action never materialised due to the backtracking from the Government in using the algorithm, and instead, opted for teacher-evaluated grading.[1370] In a hypothetical scenario where the results of the algorithm remained for students, the consequences would be detrimental to their future, and would likely cause emotional distress to many. In terms of students claiming for either a loss of chance or emotional distress (it should be noted that this would need to amount to a recognised medical condition, such as depression[1371]) caused by the algorithm, the rules of negligence would apply, and the typical regime could prove difficult to argue due to the foreseeability concept. In line with this, due to the use of an AI system, in the future, the AI Liability Directive or similar may govern the basis in proving liability.

[1365] Provisional Agreement for the AI Act (n 103) Article 6.

[1366] ibid Annex III 3(b).

[1367] ibid Article 16.

[1368] ibid Article 29.

[1369] Good Law Project, 'Legal action over A-Level results fiasco' (News, 16th August 2020) <https://goodlawproject.org/a-level-results-fiasco/> accessed 19th September 2023.

[1370] Office of Qualifications and Examinations Regulation (Ofqual), 'Statement from Roger Taylor, Chair, Ofqual' (Press Release, 17th August 2020) <https://www.gov.uk/government/news/statement-from-roger-taylor-chair-ofqual> accessed 19th September 2023.

[1371] *Hinz v Berry* [1970] 2 QB 40.

Applying the Directive, the claimant (in this case, a hypothetical student) would most likely be claiming against Ofqual (the provider of the AI system), due to either a loss of chance (in the form of a chance of further education, funding, or scholarship) or emotional distress (due to the psychological impact of such grading systems). For such a claim to be successful and the provider to be held liable, the hypothetical student would need to prove the provider's non-compliance with their obligations under the AI Act, for example, failure to comply with Article 9, a failure to identify and analyse the risks of the system, particularly in terms of the right to an education,[1372] and the risk of discrimination which has been the basis of criticism in reports on the algorithm.[1373] This basis of discrimination could have also amounted to a breach of Article 10, in the provider needing to ensure the data processed is representative, to avoid potential bias.[1374]

If Ofqual had failed to meet these standards, this would amount to a breach of care, and hence, the next element to be established by the hypothetical student would be to prove a reasonably likely link between the breach of care and the output itself, which means in context that it would need to be established that there is a likely link between the lack of analysing the risks of the system and/or the lack of mitigating bias within the system, and the grade outputted by the system. The 'reasonably likely' standard here would be sufficient for the claimant, and hence allow them to continue with the claim, by establishing that the output from the system itself gave rise to the damage, which in this scenario, would be self-explanatory. However, to prove that the output would have been different if these standards were met may be the difficult part of application, particularly if systems lack transparency.

This scenario shows the likelihood of the provider within the scenario, being liable for the harm caused, and hence, the adequacy of the rules in establishing liability. Applying the rules of negligence without the reliance on the proposed regulations would amount to issues with proving foreseeability of the damage caused by the

[1372] ECHR (n 3) Protocol 1 Article 2.

[1373] Melissa Fai, Jen Bradley and Erin Kirker, 'Lessons in 'Ethics by Design' from Britain's A Level algorithm' (Gilbert and Tobin, Lexology, 11th September 2020) <https://www.lexology.com/library/detail.aspx?g=af0fcf0c-8f56-4f8e-ab7b-c3ece59bfdf5> accessed 20th September 2023; Anthony Kelly, 'The great algorithm fiasco' (BERA Blog, 21st May 2021) <https://www.bera.ac.uk/blog/the-great-algorithm-fiasco> accessed 20th September 2023.

[1374] Provisional Agreement for the AI Act (n 103) Article 10(3).

defendant. Therefore, it can be concluded with the use of the case study, that the regulations have somewhat clarified the rules for fault-based systems and the issues that have been brought to light for negligence and tort claims.

5.3 Proposals for AI Regulation by the UK

In the UK, the current Government are making their own movements towards regulating AI, with a Policy Paper named 'establishing a pro-innovation approach to regulating AI' published in 2022.[1375] From the name of the paper itself, the objectives of the proposals are made clear and are centralised on promoting innovation, an approach that seems to overlook the ethical issues and concerns that have been raised in the area. However, this objective is clarified with the adjustment of supporting responsible innovation, highlighting the features of safety and security. The Policy Paper was published along with the new Data Protection and Digital Information Bill, forming the new era of regulation post-Brexit, although it should be highlighted that this Bill was withdrawn in early 2023, with a second Bill introduced on the 8th of March 2023.[1376]

In the AI policy paper, the UK name themselves as being at the forefront and a superpower of AI technological developments, and that the country intends to use AI for its potential growth and societal benefits.[1377] The Policy Paper sets out the intention to create regulatory frameworks that are proportionate, light-touch and forward-thinking, which they call 'essential' to battle the pacing problem and to drive innovation, clarification and confidence for businesses, and stated that a White Paper should be released later in 2022.[1378] The White Paper did not emerge until March 2023, which may reflect the UK's approach in precaution in the reaction and reflection of the criticism raised to the EU's proposals. The White Paper states that a proportionate approach will be taken, supporting innovation but also addressing risks where necessary.[1379]

[1375] UK Government (n 1129).
[1376] Data Protection and Digital Information Bill (n 925).
[1377] UK Government (n 1129).
[1378] ibid.
[1379] Department for Science, Innovation and Technology (n 99) 2.

In comparison to the EU's approach, the UK's is less centralised, focusing on leveraging existing legislative frameworks and powers, and relying on regulators to target guidance for their own domain. The approach will focus of assessing AI based on the usage, rather than the technology generally, and that five cross-sectoral principles will underpin the future adaptions to frameworks;

> "*(i) safety, security and robustness;*
> *(ii) appropriate transparency and explainability;*
> *(iii) fairness;*
> *(iv) accountability and governance;*
> *(v) contestability and redress*;"[1380]

The use of cross-sectorial principles can be compared to the GDPR's data protection principles,[1381] which lie at the heart of the regulation and can be applied across the board. The UK's cross sectoral principles are stated to build on the OECD principles on AI,[1382] and demonstrates the UK's commitment to them, which include the following: ensuring AI is used safely, ensuring AI is technically secure and functions as designed, ensuring AI is appropriately transparent and explainable, embedding considerations of fairness into systems, defining responsibility for AI governance, and to clarify route to redress or contestability.[1383]

Similarly to the EU, the UK has highlighted the difficulties in defining AI, commenting that the term itself can refer to many different systems and software, and that no single definition would be suitable for every scenario. The Government proposed the following definition in their policy paper, which they stated was sufficient for their own purposes; "*machines that perform tasks normally requiring human intelligence, especially when the machines learn from data how to do those tasks*".[1384] Alike to the EU, the UK have struggled in retaining a consistent definition. This can be seen by the differing definition of AI within the statutory instrument related to the National Security and Investment Act, as "*technology enabling the programming or training of a device or software to (i) perceive environments through the use of data; (ii) interpret data using automated processing designed to approximate cognitive abilities; and (iii) make recommendations, predictions or decisions; with a view to achieving a specific*

[1380] Artificial Intelligence (Regulation) Bill (n 101) Cl 2.
[1381] GPDR (n 1) Article 5.
[1382] OECD (n 626) 1.1-1.5.
[1383] UK Government (n 1129).
[1384] UK Government, *National AI Strategy* (September 2021) 16.

objective".[1385] The Government defend this difference in definitions by stating that clarity and preciseness is a necessity for legislative definitions,[1386] but argue that the focus on definitions is not an essential job for the state. An alternative approach is proposed, through allowing maximum flexibility and putting no boundaries on what can amount to AI, leaving the decision to be made by regulators or relevant bodies on the scope they deem fit within their field.

The Government highlight the drawbacks to this approach themselves, which is based around inconsistency and a lack of legal certainty, particularly if definitions eventually are given through the common law.[1387] This is the reasoning underpinning the Government's approach in the White Paper to base the scope of AI on core characteristics, those being adaptiveness and autonomy, informing the scope of future regulatory frameworks, but also allowing regulators to evolve more detailed definitions that are better suited to their particular sectors.[1388] This aligns with the suggestions proposed by Tech UK, who argue against a fixed definition, to allow sector-specific definitions by regulators when and as needed.[1389] The regulators being referred to here are identified within the AI Policy Paper as the ICO, Competition and Markets Authority (CMA), Ofcom, Medicine and Healthcare Regulatory Authority (MHRA) and the Equality and Human Rights Commission (EHRC). Criticism has been raised in asking multiple regulators in the future to interpret and enforce the cross-sectorial principles set out in the paper, in which the policy paper argues coordination can be ensured through introduction of platforms to ensure coherence.[1390] In response to this, the Artificial Intelligence (Regulation) Bill, introduced to the House of Lords on the 22nd November 2023 intends to create a body to be known as the 'AI Authority'; with the body to ensure the relevant regulators take into account AI systems, and ensure alignment of approaches to factor in AI systems.[1391]

The Government's approach has few similarities with the EU's AI Act, such as the focus on high risk systems, but there exist many more differences between the two jurisdictions. More generally, the UK are reluctant to impose a broad framework on AI

[1385] National Security and Investment Act 2021 (Notifiable Acquisition) (Specification of Qualifying Entities) Regulations 2021, Schedule 3(1).
[1386] UK Government (n 1389) 16.
[1387] UK Government (n 1129).
[1388] Department for Science, Innovation and Technology (n 99) 22.
[1389] TechUK, 'Government proposals for UK AI regulation' (TechUK Website, 18th July 2022) <https://www.techuk.org/resource/government-proposals-for-uk-ai-regulation.html> accessed 26th September 2023.
[1390] Oliver Yaros, Ondrej Hajda, Mark A Prinsley, Reece Randall and Ellen Hepworth, *UK Government Proposes a New Approach to Regulating Artificial Intelligence (AI)* (Mayer Brown, August 2022) 2.
[1391] Artificial Intelligence (Regulation) Bill (n 101) Cl 1.

alike to the AI Act, and are aiming for a sector-specific and de-centralised approach, meaning that although there may be a general enforcement board (AI Authority as proposed in the Bill[1392]), its basis will be to oversee other regulators, with the justification that this approach will be more adaptable to technological change. On categorising systems as unacceptable or high risk, the classification would be left to the regulators to decide on the context of systems within the specific sector, rather than a centralised list of the risk categories, alike to the Annexes of the AI Act. The UK intends to encourage regulators to consider lighter-touch options in the first instance, opposed to the EU's compulsory obligations, with the justification to avoid unnecessary burdens to innovation. From these differences, the argument can be made that the UK is focusing more so on promoting and encouraging innovation across sectors rather than on the perceived drawbacks and concerns related to the use of AI, which could leave more risk to fundamental rights and society. However, it should be noted that many systems and providers that issue AI systems within the UK would likely be subject to the AI Act due to its extraterritorial scope, and hence, the differences between the approaches may consequent in minimal need for discussion.

5.4 Proposals for AI Regulation from Other Jurisdictions

The final section of this Chapter intends to highlight regulatory approaches which are taking shape outside of the EU and UK, to assess whether inspiration can be taken from other jurisdictions in establishing the best practice to regulate AI effectively. The AI Act was the first proposed regulation of its kind worldwide, so other jurisdictions are not at the same stage of drafting and debating legislation at a similar level to the EU as of yet, but there have been some steps made towards regulation in a number of jurisdictions. This section will begin with a focused consideration of the US, looking at White House proposals, and those considered on a federal level, with a focus on California in particular. This section will also consider brief proposals from the jurisdictions of Canada and Brazil to highlight the similarities and differences between approaches and priorities in regulating the AI in the digital age.

5.4.1 The United States

In the US, there have been several reports noted by the White House of the ways in which AI can negatively affect society, with mention of privacy and discrimination

[1392] ibid Cl 1.

concerns, the risk of bias and inaccuracies, and the misuse and abuse of systems to abuse marginalised and vulnerable groups in society.[1393] The White House acknowledges the need to protect society of these risks, and acknowledge that through history, the reinterpretation and expansion of rights have been needed to keep up with progressive norms, with technological advancements posing the next stage in need of a reconsideration of rights. To address this, since 2021, the White House have backed the development of a 'Bill of Rights for an AI-powered World', stating that it is critical to ensure data-driven technologies reflect and respect the democratic values of the country.[1394]

A White Paper setting out these views was introduced in October 2022, involving practical guidance to government agencies and encouragement for technology companies, researchers, and civils societies to integrate protections into systems, working towards a human-centric design for AI.[1395] The White Paper, named the 'Blueprint for an AI Bill of Rights' chooses to define 'automated systems' rather than AI, as "*any system, software, or process that uses computation as whole or part of a system to determine outcomes, make or aid decisions, inform policy implementation, collect data or observations, or otherwise interact with individuals and/or communities.*"[1396] The full definition is much lengthier, and provides examples and techniques used within automated systems, such as ML, statistics, and AI techniques. The White Paper is based upon five non-binding principles of focus to minimise the risk of harm from the use of AI, which are:

- Safe and effective systems
- Algorithmic discrimination protections
- Data privacy
- Notice and explanation
- Human alternatives, consideration, and fallback.[1397]

[1393] White House, 'ICYMI:Wired (Opinion): Americans Need a Bill of Rights for an AI-Powered World' (White House News and Updates, 22nd October 2021) <https://www.whitehouse.gov/ostp/news-updates/2021/10/22/icymi-wired-opinion-americans-need-a-bill-of-rights-for-an-ai-powered-world/> accessed 1st October 2023.
[1394] ibid.
[1395] White House (n 1122).
[1396] ibid 10.
[1397] ibid 15-48.

The Blueprint intends to address all AI systems, rather than applying rules based on a risk-based approach with a predominant focus on high risk systems alike to the EU and UK. The introduction of this blueprint has been alongside a risk management framework, mandated by Congress and published in full by the National Institute of Standards and Technology in early 2023, providing guidance for managing risks in the development, use and evaluation of systems.[1398] The rules have some similarity to the risk management obligations included within the AI Act under Article 9,[1399] however the main difference between the two proposals is that the US draft framework is voluntary, and intends to assist in incorporating trustworthiness into AI systems rather than including obligations that underpin a compulsory framework.[1400]

The US approach is comparable to the UK in some respects, particularly in terms of introducing general principles that provide the overall scope to achieve in the use of AI and maintaining a less centralised, and more flexible approach. Alike to the UK, this would allow sectors in the future adopt legislation with interpretation of the principles in the appropriate scope, avoiding the challenges faced by the EU in regulating AI as a whole and individual concept.[1401]

The proposed Blueprint and Risk Management Framework has also been accompanied by a number of other legislative rules, focused on targeting particular areas, including; the National Defense Authorization Act 2023, which is centred on AI used by the US Department of Defense and other federal agencies,[1402] the Algorithmic Accountability Act 2022, focused on targeted impact assessments to mitigate biased decision-making,[1403] the American Data Privacy and Protection Act 2022 which looks at re-interpreting existing privacy rules in the context of AI technologies, including risk assessments,[1404] and the Health Equity and Accountability Act 2022, which sets out the use of algorithms in the healthcare

[1398] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (US Department of Commerce, January 2023).
[1399] Provisional Agreement for the AI Act (n 103) Article 9.
[1400] National Institute of Standards and Technology (n 1403) 2.
[1401] Nicol Turner Lee and Jack Malamud, 'Opportunities and blind spots in the White House's blueprint for an AI Bill of Rights' (Brookings, 19th December 2022) <https://www.brookings.edu/articles/opportunities-and-blind-spots-in-the-white-houses-blueprint-for-an-ai-bill-of-rights/> accessed 2nd October 2023.
[1402] National Defense Authorization Act for Fiscal Year 2023 (S.4543), 117th Congress (2021-2022).
[1403] Algorithmic Accountability Act of 2022 (H.R.6580), 117th Congress (2021-2022).
[1404] American Data Privacy and Protection Act (H.R. 8152), 117th Congress (2021-2022).

sector, with the aim to prevent bias in access, quality and outcomes in healthcare.[1405]

Particularly due to the US being a federal state, there is more chance of fragmented and patchwork regulation if no guidance is given from the central institutions, similar to the concerns faced by the EU if they did not step in with proposals for regulation. This has already been seen with the US, where some States have already banned FRT due to reports of bias and inaccuracies,[1406] and some States implementing more rules than others.[1407] In late 2023, President Biden issued an Executive Order on Safe, Secure and Trustworthy Artificial Intelligence, labelled by the White House itself as 'landmark', and 'ensuring that America leads the way' in managing the risks of AI.[1408] The Order intends to show a government-wide effort to guide the responsible development and deployment of AI, building on the work in the Blueprint for an AI Bill of Rights, and identifies eight overarching policy areas of focus, which majorly align with the principles set out in the Blueprint.[1409] It is hoped that the Order provides consistency and oversight to State regulations.

5.4.2 The US State of California

With particular focus on California, recent proposals have been made towards an establishment for an Office of AI, contributing to the State's overall efforts to regulate the use of AI.[1410] The proposed Office intends to guide the design, use and deployment of automated systems, to ensure that such systems comply with federal regulations, and to ensure that bias in systems is minimised.[1411] The focus on mitigating bias, in addition to ensuring transparency and the protection of privacy and civil liberties aligns with the Blueprint proposed by the White House, however,

[1405] Health Equity and Accountability Act of 2022 (H.R. 7585), 117th Congress (2021-2022).
[1406] E. Barlow Keener, 'Facial Recognition: A New Trend in State Regulation' (Womble Bond Dickinson Blog, 29th April 2022) <https://www.womblebonddickinson.com/us/insights/alerts/facial-recognition-new-trend-state-regulation> accessed 2nd October 2023.
[1407] Ban Facial Recognition, 'Ban Facial Recognition' (Website, Interactive Map) <https://www.banfacialrecognition.com/map/> accessed 2nd October 2023.
[1408] White House Executive Order (n 1122).
[1409] White House (n 1122); Congressional Research Service, *Highlights of the 2023 Executive Order on Artificial Intelligence for Congress* (November 2023) 1-2.
[1410] Ayesha Gulley and Airlie Hilliard, 'The Proposed Amendments to California's Employment Legislation Regarding Automated-Decision Systems' (Holistic AI, 18th June 2023) <https://www.holisticai.com/blog/california-employment-legislation-proposed-amendments> accessed 2nd November 2023; Assembly Bill No. 1651 (n 1122).
[1411] An act to add Chapter 5.9 (commencing with Section 11549.80) to Part 1 of Division 3 of Title 2 of, the Government Code, relating to state government (Senate Bill No. 313) 2023.

concern has been raised in relation to the enforcement mechanisms to ensure compliance.[1412] California had a more focused approach in terms of regulation, with proposed amendments to employment regulations to address bias and discrimination,[1413] and an introduction of a proposed Workplace Technology Accountability Act,[1414] restricting the collection of data of employees, requiring data protection impact assessments, which could be seen as similar to the GDPR requirements,[1415] in addition to algorithmic impact assessments. Unfortunately, however, this Act did not surpass the legislative process, and will not be progressing into law.

In comparing this approach to the EU, California was clearly taking a more targeted approach looking at the employment sector in particular, whereas the AI Act is much more expansive, covering all sectors that use AI. The approaches are similar in terms of the assessments proposed to be carried out, with California's approach having comparisons with the conformity assessment included within Article 43.[1416] In California's failed Act, it is specified clearly that such conformity assessments must be carried out by a third party, something that is arguably amiss and lacking clarity within the AI Act. If California had continued this approach for other sectors, it could arguably surpass the rules set out in the EU, with targeted provisions for the relevant sector, rather than a general approach to all systems.

5.4.3 Canada

In Canada, a focus has been made on the use of personal data related to AI tools, through a proposed Digital Charter Implementation Act,[1417] playing a part in the advancement of Canada's 'digital charter', intended to strengthen privacy protections and guide the advancements of the digital economy.[1418] The Digital Charter is based upon ten principles: universal access; safety and security; control and consent; transparency, portability and interoperability; open and modern digital government; a

[1412] ibid.
[1413] Gulley and Hilliard (n 1415).
[1414] Assembly Bill No. 1651 (n 1122).
[1415] GDPR (n 1) Article 35.
[1416] Provisional Agreement for the AI Act (n 103) Article 43.
[1417] Digital Charter Implementation Act (n 1122).
[1418] Innovation, Science and Economic Development Canada, *Canada's Digital Charther in Action: A Plan by Canadians, for Canadians* (2019).

level playing field; data and digital for good; strong democracy; freedom from hate and violent extremism and strong enforcement and real accountability.[1419] The principle based approach taken within this Charter has clear comparisons to the UK's cross-sectorial principles and the principles set out in the Blueprint, showing the priorities when it comes to regulating AI.

The proposed Digital Charter Implementation Act would lead to further implementation of legislation, such as a Consumer Privacy Protection Act, which has some similarities with the GDPR, but is predominantly focused on the use of AI.[1420] The Act intends to focus on a human-centric approach to controlling data, with specific data such as the personal information of minors referred to as sensitive data.[1421] This approach is one that aligns with the thesis' argument, taking inspiration from the GDPR rules but with a refocus on AI products and systems to better align regulation. The approach within this proposal is clearly following the GDPR approach in other ways too, as penalties for non-compliance would result in fines of 5% of annual turnover or $25 million, whichever is greater.[1422] This approach with penalties has been reflected in both the GDPR and AI Act, giving severe repercussions to those who do not comply with the rules. This proposed Act, however, has a striking difference to other approaches discussed so far, as it imposes criminal penalties for those who misuse emerging technologies, and who fail to follow the rules on the responsible development of AI systems.[1423] This unusual penalty could lead to interesting case-law, in which sanctions could be much more severe than just financial.

This approach is replicated in an additional proposed Bill, the Artificial Intelligence and Data Act, which would require businesses to comply with a number of requirements, including documenting their reasoning for using AI systems, and integrating safeguards.[1424] This proposed Act would also impose criminal sanctions on businesses who unlawfully obtain data with the intention to cause serious harm or economic loss, in addition to criminal sanctions for creating an AI system with

[1419] ibid 15.
[1420] Bill C-27 (n 1122).
[1421] ibid.
[1422] ibid s128(a).
[1423] ibid s38-40.
[1424] ibid; Government of Canada, *The Artificial Intelligence and Data Act – Companion Document* (2022).

knowledge that it could cause serious harm, either physical or psychological, or property damage.[1425] The rationale behind the criminal offences under the proposed Act is to prohibit and punish AI-related activities by those who have the knowledge, or who can appreciate the harm they are causing or at risk of causing.[1426] Re-emphasising the point above, it will be interesting to see the subsequent case-law that stems from this proposed Act if enforced, particularly in terms of how the courts will assess what amounts to knowledge that a system *could* cause serious harm or property damage.

Whilst the above Bills are still awaiting Parliamentary approval, Canada has chosen to fill the gaps by issuing a number of non-binding guidelines, including the Guide on the use of generative AI for the use of AI in the public sector,[1427] and a Voluntary Code[1428] which sets out a number of ethical safeguards to be upheld, including accountability, safety, fairness, transparency and human oversight. These non-binding initiatives reflect the advantages of the use of soft law filling gaps whilst hard legislative development takes place.

5.4.4 Brazil

In Brazil, a proposed Bill titled (as translated in English) as the Artificial Intelligence Bill,[1429] and alike to many others discussed, takes a principled-based approach, focusing on the safety and ethical aspects of systems including: beneficial purposes of systems; human-centric approaches; non-discrimination; pursuit of neutrality and transparency. [1430] The Bill's aims are set out within Article 3, and include promoting sustainable and economic development of the well-being of society, in addition to increasing the competitiveness and productivity of the use of AI.[1431] The Bill contrasts the others in the field discussed so far due to its shorter length, comprising of only ten articles, and focuses on establishing centralised principles that should be

[1425] ibid.
[1426] ibid.
[1427] Government of Canada, 'Guide on the use of generative artificial intelligence' (Canadian Gov Website) <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html> accessed 4th September 2024.
[1428] Government of Canada, 'Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems' (2023, Canadian Gov Website) <https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems> accessed 4th September 2024.
[1429] Artificial Intelligence Bill, No. 21 of 2020 (n 1122).
[1430] ibid Article 5.
[1431] ibid Article 3.

followed, rather than a full regulation of the use and deployment of systems alike to the AI Act. Within these articles however, there is inclusion of risk-based management, similar to the approach within Article 9 of the AI Act.[1432] The Brazilian Bill purposely detracts from following the AI Act, where the rapporteur has reportedly stated that definitions and classification based on the supposed risk of systems will be left to subsequent sectoral regulation or self-regulation,[1433] which departs from the centralised approach seen within the US and EU.

The Bill has been subject to criticism on this basis, where it has been highlighted that the Bill may equate to weakened civil liability and data protection, with a focus on Article 6 in particular.[1434] Article 6 of the Bill governs the proposed rules for civil liability, whereby in fault-based claims, an individual who suffers harm would need to prove that there was a mistake committed by a provider or deployer during the machine's lifecycle and that the provider or deployer was either at fault or negligent.[1435] This regime would put a distinctive burden on the claimant who has suffered harm, making it difficult to be adequately compensated, and departs from the reduced burden suggested within the AI Liability Directive.

5.5 Conclusion

Through evaluation of the current regulatory landscape in the jurisdictions discussed, it is evident that the EU's AI Act goes much further in regulating AI than other nations and regions, giving the likelihood of worldwide domination of rules alike to when the GDPR was first implemented. Although the EU's AI Act is subject to its own criticism, including its broad categorisation of systems under the risk-based approach, it is also the most detailed approach, in that rules and obligations are imposed on providers and deployers in a mostly clear manner. The Act encourages the

[1432] Provisional Agreement for the AI Act (n 103) Article 9.

[1433] Câmara dos Deputados, 'Chamber approves project that regulates the use of artificial intelligence' (Federal Legislative Body, News, Translated, 29th September 2021) <https://www.camara.leg.br/noticias/811702-camara-aprova-projeto-que-regulamenta-uso-da-inteligencia-artificial?utm_source=POLITICO.EU&utm_campaign=25c6120bdd-EMAIL_CAMPAIGN_2021_11_17_09_59&utm_medium=email&utm_term=0_10959edeb5-25c6120bdd-190866048> accessed 4th November 2023.

[1434] José Renato Laranjeira de Pereira and Thiago Guimarães Moraes, 'Promoting irresponsible AI: lessons from a Brazilian bill' (Heinrich-Böll-Stiftung- German political foundation blog, 14th February 2022) <https://eu.boell.org/en/2022/02/14/promoting-irresponsible-ai-lessons-brazilian-bill> accessed 5th November 2023.

[1435] Artificial Intelligence Bill, No. 21 of 2020 (n 1122) Article 6(6).

safeguarding of fundamental rights and sets out explicit obligations and penalties, offering a strong starting point to encourage responsible use of AI.

Also, the proposed AI Liability Directive sets out commendable redress mechanisms for fault-based claims, in which the reduction of the burden of proof on claimants will ensure adequate compensation can be accessed if systems do act in a way that results in harm. This proposal does well to balance innovation with accountability, offering individuals clear pathways to damages in scenarios where AI systems produce harmful decisions. Such provisions reinforce the importance of the rule of law and access to justice, which is fundamental to promoting public trust in AI.

In contrast, other jurisdictions, such as the UK and US have moved away from a risk-based approach, and instead have focused on a more flexible, principle-based approach. This gives room for subsequent sector-specific regulation to be made, which could allow more specific rules for the use of AI within the context of those sectors. Whilst this approach offers adaptability, it also risks fragmentation and inconsistency between sectors, making it more difficult to implement a framework which can combat the complexities of AI systems in a uniform way. The principle-based approach, although appealing, may lack the enforceability mechanisms and clarity that the EU’s risk-based approach offers.

Whilst the EU’s AI Act sets a strong precedent for comprehensive AI regulation, more flexible approaches could complement this model through providing valuable lessons in balancing regulation with innovation. To continue the discussion, the thesis will conclude in Chapter Six by connecting the discussion in previous chapters together with the discussion featured in this chapter on the current regulatory landscape, to reinforce the thesis’ objectives and address the main aim of proposing further recommendations for the promoting and encouraging the ethical use of human rights centred AI. Through this discussion, the latter half of the third research question will be met, addressing further solutions to be proposed to strengthen safeguards and encourage ethical innovation further.

Chapter 6: Conclusion

6.1 Summary

Through consideration of the legal, ethical, and technological issues that have arisen in the wake of AI, the prior chapters have identified emerging themes in the literature, evaluated the current shortcomings in the law, and assessed the complexities in regulating the technology, to give the key findings and recommendations in this concluding chapter. Through the presentation of solutions and recommendations to promote ethical AI within a human right centred framework, this thesis will address the research aim by producing up-to-date, justified, and well-researched solutions towards human rights based regulation for AI. In support of the hypothesis that led this research, the thesis advocates that inspiration needs to be taken from the GDPR in the creation of a new framework, not only to allow alignment to data protection, but also to achieve a similar impact to the new digital world.

With consideration to the recommendations and research findings set out in the next section of this conclusion, the thesis addresses the overarching research question; 'to what extent should AI be regulated to ensure an ethical framework centred on human rights.' This thesis has focused on the key regulatory aspects of AI, examining the type of regulation that is needed, and what the foundation of such a framework should include. Being one of the first blanket regulations for AI in the world, this project has used the EU's AI Act as a prominent example, and has identified the strengths and shortfalls of the EU's Act to aspire to greater levels of human rights protection and safeguarding here in the UK. Given the freedom Parliament now has to depart from EU legislation, this thesis stands as a recommendation to policymakers and regulators on the lessons that can be learned from the EU regulatory process for making a widespread AI framework.

Alike to the EU, this thesis supports the view that a new framework for AI must be introduced to properly address the conflict in the law, and human rights concerns. Central to the thesis' arguments it the proposal for a new AI regulatory framework which draws inspiration from the GDPR, not only to ensure data protection and complement the DPA, but also to create a robust, ethical foundation for AI regulation. The thesis advocates that a new framework should seek alignment with the GDPR principles, rather than cause conflict or confusion in the alignment. However, it also argues that this framework must extend beyond just data protection to ensure the unique aspects of AI are regulated effectively, capturing and addressing the issues of foreseeability, accountability, and accuracy, fairness and bias. It is key that future regulation focuses on the protection of human rights and provide safeguards to society, whilst also being proportionate, and ensuring that the benefits of AI and the technological transformation it brings can be reaped.

The conclusion continues by presenting the core research recommendations and findings through the reiteration of the hypothesis, and emphasis on how the research objectives have been achieved, and the research questions answered. Following the research findings, which include the development of a human rights centred framework, alignment with the DPA, and enhanced accountability mechanisms, the conclusion revisits the significance of the research, and the contribution it makes to the field and wider policymaking. The conclusion also considers the limitations of the research, and thoughts on the future, finishing by revisiting the key points and findings made.

6.2 Recommendations for a Future Human Rights Centred Framework for AI

As stated, to ensure efficient regulation of AI, and to provide appropriate safeguards to society, binding legislation must be introduced. Due to England and Wales not currently having a focused framework for AI, one must be implemented to allow future rules and regulations to build upon this. To solidify a strong stance post-Brexit, England and Wales should not fall behind the regulatory developments that are already taking place within the EU, and given the rise of the use of AI technology, introducing a new legislative framework on AI must be a priority. This section combines the suggestions presented in the thesis, to offer a comprehensive solution to regulating AI, offering an opportunity for England and Wales to set themselves apart from the EU, and the advantage of being able to learn from the EU's mistakes.

6.2.1 Human Rights Focus

Introducing a new regulation allows the chance for Parliament to make clear its intentions regarding the topic area of focus. In this context, a future framework must make clear that it intends to uphold human rights standards, provide safeguards to society, and work to promote ethical AI. Society's views of AI have been led by futuristic film and entertainment, and media attention has been focused on the unethical use of AI, which has led to a decline in public trust. By providing clear, and concise human rights centred regulation, society would see the efforts made to ensure that the benefits of technology can be reaped, but also, the priority given to the safeguarding of their livelihoods and rights. A suggested purpose to be contained within a future framework would be to uphold human rights protections and provide safeguards for developing the proportionate use of ethical AI.

6.2.2 Inspiration from the GDPR

As proposed throughout this thesis, a new framework for AI could complement, and work in collaboration with existing legislation, such as the GDPR, to add clarity for data controllers, and strengthen protections for data subjects. A future framework for AI having similarities to data protection legislation would emphasise to parties involved the importance of the regulation, and also offer confidence in the ability to comply with the obligations within it. This thesis has focused on AI technology that uses datasets that comprise of personal data or processes personal data to decision-make, which provides further justification to take inspiration from the GDPR to maintain alignment. Although this could provide the focus for future regulation, the principles set out should apply to all AI, but be mandatory for those that use personal data.

The definition that sets the basis of personal data in the GDPR is already understood, so data controllers would be able to apply this term with ease, due to their other data protection obligations. The definition of personal data is broad, and gives a non-exhaustive list of examples. A similar standard could be met for AI, and to remain consistent, the proposed definition reiterates the one given within the introduction, that AI includes products or components that have differing degrees of intelligence and autonomy, and receive data inputs to produce outputs, and can perform one or more specific human tasks.

Further terminology (such as 'data controllers' and 'data processors'), the data protection principles, and enforceability mechanisms can also be drawn from the

GDPR and repurposed into the basis for future AI regulation. Creating a regulation by focusing on its alignment with already existing rules would allow for easier implementation, and better acceptance from all parties involved, highlighting the similarities between the standards set.

6.2.3 AI Data Protection Principles

Having the main provisions within a future framework apply to all AI systems that include personal data in their datasets or process personal data would ensure a consistent standard of obligations, regardless of the risk posed. If personal data is concerned in decision-making or the use of an AI system, society is involved, and hence obligations would provide for a human rights and societal safeguarding approach, rather than a risk-based approach, reflecting the difference between these suggestions, and those introduced in the EU. Taking inspiration from the GDPR, it is suggested that AI Data Protection Principles are introduced to set the basis of future regulation, and give the starting point for compliance. Given the focus, the data protection principles would apply in circumstances where personal data makes up the dataset, so these principles are suggested to work in addition to, and not instead of, those within data protection regulation. These data protection principles are proposed to consist of the following:

Transparency, Explainability, and Interpretability

To ensure the ability to offer redress to individuals affected by decisions, the components of transparency, explainability, and interpretability are essential. These principles would give the starting point for compliance. The principle of transparency would entail an obligation on manufacturers of AI, to ensure the workings of systems could be understood, and for guidance to be provided for those who deploy systems to understand the workings behind decisions. The principle of explainability would provide the right to an explanation of a past decision, through offering counterfactual explanations. Counterfactual explanations would allow for interpretability, and ensure laymen would understand the factors considered in decision-making. Where systems cannot be transparent (due to technical limitations), manufacturers would need to apply for an exception, which would be accepted by an oversight authority where it can be shown that the other suggested principles have been upheld and are subject to heightened testing requirements.

Lawfulness

It can only be expected that every possible use of AI would not fulfil the purpose of achieving ethical AI. For this reason, it should be required that for systems to be available on the market and make decisions using personal data, an authorisation certificate must be required. Such a certificate is proposed to be given by an established new authority, detailed below. Practically, it would be unfeasible to order a ban on all systems currently used until such a certificate could be obtained, and for this reason, it is suggested that there be an 18-month grace period to obtain a certificate. If one has not been obtained by the deadline, it would be unlawful to have the system available on the market, for both public and private bodies.

At first glance, an 18-month grace period may be seen as too long, but this would allow developers and deployers to collate the evidence needed for such a certification and allow for the establishment of a new authority and the capacity to deal with a certification scheme, without unnecessary backlog. It should also be noted, that those systems that impose a disproportionate level of harm to society should be banned, and an oversight authority could aid in drawing up such a list, and amend it for future purposes.

Non-Discrimination

Those who use AI, by which the datasets include personal data or process personal data to decision-make, would need to comply with the principle of non-discrimination. This would be focused on mitigating bias that falls within the scope of the protected characteristics accounted for under the Equality Act,[1436] and aim to address the concerns and countless reports of bias infiltrating systems. To comply with this principle, evidence would have to be gathered during the pre-market phase to show the efforts made to detect and correct any evidence of bias. In addition to this, this principle will apply post-market, and evidence of bias would need to be addressed within a timely manner set by an oversight authority.

[1436] Equality Act 2010, s4.

Accuracy

To comply with this principle, again, there would be a focus on pre-market and post-market obligations. If issues with accuracy arose through these obligations, the principle would require deployers to address and evidence improvements within a timely manner set by an oversight authority. The thesis has noted the difficulties in setting an exact standard to be expected, and for this reason, argues that accuracy must be at a higher standard to an average human counterpart in the field as a starting point. However, monitoring requirements would focus on the number of false positives and false negatives, and ensure that accuracy is measured regularly, and consistently.

Fairness

This principle would intend to fill the gaps left by the above and would serve as a requirement when systems show signs of unfairness through the monitoring requirements, or where a decision has been identified as unfair. As above, where an unfair decision has been identified, deployers must act to correct and improve systems promptly, within a time limit set by an authority. Due to the subjectiveness of 'fairness', and where a decision is contested, a branch within the newly established body will work to determine the threshold based on the decision itself, and will apply penalties based on their decision.

6.2.4 Pre-Market Monitoring Requirements

Before a system can be officially permitted on the market, the system must gain certification. This would apply to all systems, not just those who use personal data. To gain certification, several proposed steps must be complied with, at which point the authority carries out an independent audit to test the evidence collated. The steps to achieve before submitting for certification are proposed to include mandatory FRIAs and evidence collation, and are discussed further below.

The certification system has been proposed within the AI Act under Article 49,[1437] however, the suggested certification procedure would need regular oversight from an authority, and would expire if documentation is not updated or reviewed at least

[1437] Provisional Agreement for the AI Act (n 103) Article 49.

annually. The suggested certification system would also be monitored to help aid with redress and enforcement mechanisms, which are discussed in the next sections (6.2.5 - 6.2.6).

A Fundamental Rights Impact Assessment

Requiring data controllers to complete a FRIA would complement the process within the GDPR, whereby a DPIA must be carried out. A FRIA would force data controllers and processors to take more consideration into the societal impact of the AI systems they use, and would require justification for their use, in addition to a detailed assessment of what has been done to safeguard against infringement of rights. The FRIA would provide a scope of focus on the Convention rights,[1438] but would allow for further rights to be considered depending on the context of use, and be centred around the concept of proportionality.

The FRIA would also ask for evidence of compliance with the AI data protection principles proposed above. This would be reviewed by an oversight authority, where suggestions for improvement, or consideration of other relevant rights that are needed could be highlighted. The FRIA should focus on the use of systems, but where the use could be unclear (for example, in generative AI), all possible outcomes should be considered. This may seem unreasonable, but would force developers to limit the availability of services offered with AI if they cannot deem it as justifiable in a fundamental rights context, and in line with the AI data protection principles.

This suggestion goes further than the EU’s introduction of FRIA in Article 29a,[1439] as the foundation of the suggested approach would not be based on risk-factor, nor would this obligation only apply to systems that offer public services. This suggestion would apply to all products and services, with heightened considerations for systems that process personal data, to align with DPIAs.

[1438] Human Rights Act 1998, Articles 2-14.

[1439] Provisional Agreement for the AI Act (n 103) Article 29a.

Evidence Collation

To comply with the above, and to ensure the below obligations can be upheld, organisations would need to establish AI-specific departments and roles with teams who have the competency, and expertise to work with and analyse machines sufficiently. Organisations would also need to ensure digital literacy skills within their employees, so that documentation can be uploaded and shared with the relevant bodies correctly, and promptly. Before a system is permitted on the market, it must be thoroughly tested to ensure that the AI data protection principles and other safeguards (including data protection and human rights standards) are upheld. The datasets used for training systems are proposed to undergo a testing audit by the Oversight Board before certification is granted, and the Board can release details of what an audit would entail to allow for preparation.

6.2.5 Post-Market Monitoring Requirements

In addition to pre-market requirements, for the reasons discussed throughout the literature, it is just as important to set post-monitoring requirements once a system has gained certification and is operable on the market. Again, this would apply to all systems, rather than just those that process personal data. To make clear the differentiation between this suggestion and Article 61 of the AI Act,[1440] these obligations would apply to all systems, with heightened rules for those that process personal data, as explained further below.

Human Intervention

This requirement reiterates the key suggestions made in Chapter Four, and focuses on the alignment with the GDPR, specifically Article 22. The standard suggested for human intervention would integrate key individuals within corporations who have the role of monitoring, and overriding decisions where necessary. All decisions monitored by the suggested Intervention Officer should be logged, with logs uploaded to a database, which is shared with the overarching AI authority for evidence and review purposes. Logs should be documented whether or not an Intervention Officer chooses to act, and this should be justified, so that the responses can be reviewed internally and externally, to reduce the risk of

[1440] ibid Article 61.

confirmation bias or superficial monitoring. Where unpredictable or unexplainable decisions are identified, these should be flagged, for action both in-house and for notification on the Oversight Board for review.

Also, as discussed in Chapter Four, to strike the balance between safeguarding society and avoiding the stifling of innovation, developers who believe that the scope of the refined Article 22 should not apply to their use of AI can appeal for an exemption to the Oversight Board. To provide additional safeguards, it is suggested that all black box systems should be required to comply with the provisions of the refined Article 22, regardless of the risk associated with decisions made. However, it should be noted that this needs to be accompanied with the principle of explainability in the form of counterfactual explanations, so that an Intervention Officer can review decisions properly. Regardless of the ability to give counterfactual explanations, black box systems would still be subject to insurance, discussed further in the below section (6.2.7).

Internal Annual Audits

To renew certification, data controllers should be required to complete an internal audit, which is similar to the external audit completed before certification. This would require data controllers and processors to review and update the FRIA, complete an audit using guidelines set by the Oversight Body, and upload the results to the documentation system. Sharing these audits would account for documentation and evidence purposes, whilst also allowing the Oversight Board to identify any growing concerns in the field. If audits were insufficient, or not completed, the certification could be revoked, and data controllers would violate the lawfulness principle, and hence, be subject to penalties.

External Audits

In addition to internal audits, a branch of the Oversight Board could provide an external auditing service once systems are on the market. This could work like the Office for Standards in Education (OFSTED) system in education,[1441] where random

[1441] Ofsted, ‘About Us’ (Government Website) <https://www.gov.uk/government/organisations/ofsted/about> accessed 15th June 2023.

audits are taken within a set amount of time from certification. Once an audit has been completed, a report can be produced, with a score assigned to that system and data controller. This could promote competition, and highlight good practice for safeguarding rights, whereby others could read and learn from reports. If a low score is given, data controllers would be given a time limit to correct their mistakes before certification is revoked.

6.2.6 Enforceability

The above suggestions refer to the assistance of an external authority, suggested within the thesis to be titled the 'AI Human Oversight Board'. A new board is suggested rather than relying on authorities that already exist, such as the ICO. The authorities could, however, work collaboratively to share best practices, due to the assurance needed to safeguard human rights and society. As stated, the AI Human Oversight Board could consist of several branches, and aid in:

- granting certification;
- completing pre-certification audits;
- completing post-market audits;
- granting exemptions to human intervention;
- setting rules for best practice;
- setting further standards and giving interpretation to regulatory rules;

In addition to the above administrative and auditing duties, the Oversight Board would also be given the responsibility of enforcement. The certification scheme could provide a basis for this, where a breach of the rules could result in revocation of the certificate, and the system no longer have a lawful basis. For instances where systems are not working as intended, for example, producing several unpredictable or unexplainable results, even where the obligations are adhered to, the Oversight Board could work in conjunction with the data controller to try and identify the error, and could suspend licenses until the issue is corrected.

When an issue is identified (either through audits or through oversight mechanisms), and for example, documentation is uploaded late, or incomplete, a one-strike system could be put in place, to allow for the Oversight Board to guide those attempting to comply with the rules, before certification is removed. This would avoid the unnecessary stifling of innovation, and aim to help strike the balance needed for effective regulation. Particularly in the early stages once regulation has been introduced, this would allow for data controllers to 'get used to' and understand the extent of the rules, but after a set period, a strike system could be removed.

This system has similarities to the database suggested within Article 60 of the AI Act, however, the foundations of the obligations are proposed to apply to all products and services, with heightened rules for systems that process personal data. This removes the issues related to risk categorisation and provides consistent obligations for all systems, with particular focus on those systems that process personal data- to ensure a human-centric foundation is achieved.

It also needs to be considered how an authority could be created and structured. This could involve a legislative process, replicating the proposals within the Artificial Intelligence (Regulation) Bill,[1442] giving a statutory footing to oversight. In terms of structure, the Board could consist of several branches, to ensure domain-specific expertise is incorporated into the certification and auditing processes. The body could work in collaboration with other established existing bodies, such as the ICO, Ofcom and other regulators to ensure consistency, minimise overlaps, and avoid conflicts in guidance.

The precise roles of individuals within the Board could include certification offers, who could have the responsibility for assessing systems for compliance with the proposed regulatory requirements; audit officers, who could monitor systems and respond to issues flagged during operation; and enforcement officers, who could oversee disciplinary actions, such as the revocation of certificates and the imposing of corrective measures.

Whilst the Board is proposed to focus primarily on AI systems, it is important to consider its overlap with other bodies, such as the ICO, particularly when considering AI systems that process personal data. To avoid unnecessary complexities and duplication of regulation, a clear list of responsibilities could be

[1442] Artificial Intelligence (Regulation) Bill (n 101).

established for the Board. This could be for example, where data protection specific oversight (under the DPA) remains with the ICO as the predominant regulator, whereas the Oversight Board handles the broader issues of AI governance, including explainability, fairness, and operational compliance.

6.2.7 Accountability

For regulation to work effectively, accountability and redress mechanisms must be in place to ensure those subject to unfair or incorrect decisions can be compensated. To reiterate the suggestions made in Chapter Three, this could work through combining the ideas presented within the literature.

Firstly, it is proposed that regulation sets a duty of care standard for AI developers and deployers, to comply with the AI data protection principles, pre-market and post-market obligations discussed. If these are not met (and not identified), and consequently lead to an unfair or unpredictable decision, the developers of systems would carry the burden of responsibility, and would be strictly liable, without the requirement of having to prove causation between the breach of care and the unfair decision made. This strict liability mechanism would reiterate the importance of safeguarding society. If the obligations are not complied with, and this is identified by the Oversight Board or another authority, a fine would be put in place. Depending on the type of breach, this could also lead to the revoking of certification, emphasising the importance of adhering to the principles and obligations.

For instances where duties of care have been complied with, yet inaccurate, unfair, or biased decisions are made, a compensation scheme should be put in place to reimburse the claimant concerning the decision made, and the amount to be decided by the courts. The key concern with this suggestion is where the capital would come from. It is suggested that this is collected through several mechanisms, including a portion of the Government AI budget, collection of fines on developers who breach the rules, and money gained through a licensing scheme, which is proposed to be mandatory for systems regarded as ‘black-box’.

Black box systems would find it difficult to comply with the first AI data protection principle and therefore, other mechanisms need to be in place to ensure accountability. Due to the inability to understand systems and their decision-making, and the issues this can cause with redress, all black box systems are proposed to be licensed before certification is granted. As suggested by Karnow,[1443] licensing quotes could be based on a range of factors, including the probable risk of the system and the scale of its use.

If one of the possible duties of care is not adhered to, such a license could include a clause that the compensation scheme does not apply, and hence, responsibility would be on the developer or deployer (depending on the duty breached). If all possible duties were complied with, the scheme could become available. Before such a compensation scheme is relied on, a burden of proof could be placed on the developer or deployer to show that the decision made is comparable to other decisions made (and therefore would not have been identified as an anomaly), to relinquish their liability, and to allow the compensation fund to be used. This would avoid abusing the compensation scheme, whilst also offering fairness to those who wish to develop and deploy black box systems.

6.2.8 Funding and Education

For regulation to work effectively, and to encourage the integration and coexistence of AI in society, Parliament must give due regard to the importance and need for funding and education. For example, the monetary cost of complying with the above obligations should be considered; to comply with these rules, organisations would need to apply more skilled staff, offer training to existing staff, and allow for capacity for administrative and auditing responsibilities. Due to this, the Government could offer grants, particularly to small and medium sized enterprises to increase competition in the field.

It must also be noted that society needs to be upskilled to allow for AI to achieve the potential benefits it offers. This could form within education, from primary education to higher education, where a focus on integrating AI literacy skills is supported nationally. In addition to this, budgets should account for, and where possible, grants should be given to enable the upskilling of AI within education. Due to the pace of AI

[1443] Karnow (n 376).

growth, AI literacy needs to be embedded into everyday life, rather than just education. Due to the proposed obligations, it is important to note the growth of AI-competent jobs being introduced in the near future, which, again reinforces the support needed from the Government to prepare society for this. This is arguably more a governmental responsibility than a regulatory responsibility, but is an important point to consider when thinking about introducing regulation.

6.3 Research Findings

6.3.1 Hypothesis

To recap, the hypothesis that underpins the discussion throughout this thesis argues for the GDPR to be used as inspiration and as a starting point for a future human rights framework for AI. Throughout this thesis, similarities have been drawn to regulators and academics in the field who intentionally share this sentiment, including Spyridaki, the Chief Privacy Strategist of SAS Europe,[1444] the UN in their report on the promotion and protection of the right to freedom of opinion and expression,[1445] and those who perhaps unintentionally share the sentiment, such as within Canada's Consumer Privacy Protection Act.[1446]

Through completing the research, the hypothesis still stands strong and remains supported by the thesis; this can be seen through the proposals suggested in the above section, where fundamental aspects, such as core principles (like the data protection principles), are suggested to provide the basis of regulation. Alike to data subject rights, a new framework must ensure that it is clear to end-users the extent of their rights, and what options they have if these are not adhered to by developers or deployers of AI systems. There should also exist a clear basis for using AI systems, alike to the lawful basis of processing within the GDPR, to ensure that a balance is struck between promoting innovation, and restricting the use of systems where there is the possibility of unlawful and disproportionate harm to human and fundamental rights.

[1444] Spyridaki (n 84).
[1445] United Nations (n 82) [45] and [63].
[1446] Bill C-27 (n 1122).

The impact made by the GDPR should not be ignored, and its significance should be used as an inspiration for future frameworks, particularly in similar topic areas that transmit in the real world and the digital world online, such as the use of AI. Due to this, the argument that underpins this research reiterates this approach, that a new human rights centred framework for AI should take inspiration from the GDPR to give a foundation to a future framework.

6.3.2 Research Objectives

This section intends to confirm and reiterate the addressing of the research objectives stated in the introduction of this thesis. Through revisiting the research objectives, the purpose of this section serves to connect the research objectives to the research questions, and recap the key themes captured in the previous chapters.

Research Objective 1: To critically evaluate the ethical, legal, and technological issues posed by AI, to propose how regulation would best address these challenges.

This first objective was used as the basis of discussion within the literature review, which took a thematic approach to identify and highlight the key debates surrounding the ethical, legal, and technological issues related to AI. Through examination of the literature, the thesis took its form of focusing on the predominant aspects identified, including human rights protection, complexities with liability, and alignment issues with data protection, by which these key points identified link to RQ1 and RQ2.

Research Objective 2: To assess the human right implications caused by AI, to develop new regulatory safeguards that effectively protect society and their rights.

The second objective provided the basis of the first half of the discussion within Chapter Three, which consisted of an examination of the human right implications and impact made by AI, focusing on the issues of **accuracy**, **fairness,** and **bias** to tackle the area of discrimination, and the related impact to human rights. This discussion also linked to **transparency**, **explainability**, and **interpretability** in so far that these elements are a necessity to have the ability to measure the human rights impact made by systems. Through achieving this objective, the research presents the answer to RQ1, discussed in further detail below.

Research Objective 3: To examine current liability regimes and demonstrate the legal uncertainties they face with AI, to propose solutions to address this lack of clarity.

The third objective provided the scope of discussion for the latter half of Chapter Three, by which it was agreed that there were clear complexities and conflicts with the current application of liability regimes in circumstances using AI. This chapter built upon the discussion in the literature review in considering the most appropriate option of liability for AI cases, which included a discussion of negligence, the duties of care, and the issue of foreseeability. Through examining these areas, the discussion contributed to filling the gap in the research of how such issues could be addressed in practice, and how future regulation could reduce and eradicate the conflicts with the current liability provisions. This objective also ties into the latter part of RQ1.

Research Objective 4: To critically analyse the impact on data protection through the lens of the GDPR, and to propose potential areas of reform to strengthen alignment between AI systems and data protection rules.

The final objective provided the basis of discussion for Chapter Four of the thesis, which included examining the data protection regulations. The focus of the Chapter was on the GDPR,[1447] but the discussion also remains relevant to the England and Wales jurisdiction due to the implementation, and current stability, of the DPA.[1448] To ensure the longevity of this discussion, the Chapter also considered the intended changes that could eventually arise in the form of the Data Protection and Digital Information Bill (No2),[1449] contributing to knowledge by ensuring the discussion is up-to-date in terms of recent developments. It should be noted that the Bill is only in the early stages of the legislative process, justifying the main focus on the provisions

[1447] GDPR (n 1).
[1448] Data Protection Act 2018.
[1449] Data Protection and Digital Information Bill (n 925).

currently enacted. This research objective, alike to the others, coincides with the research questions, specifically RQ2.

6.3.3 – Research Question 1

RQ1: *How do AI systems challenge ethics, human rights, and current liability rules, and how would a new framework address these issues?*

As emphasised above, and stated throughout this thesis, the impact AI has on human rights should not be understated. To ensure a society which can be safeguarded and coexist effectively with AI, legislation must be human rights centred. The thesis has shown how AI can challenge ethics, human rights standards, and current liability regimes, and regulatory action must be achieved to strike the balance of promoting ethical AI, whilst avoiding unnecessary stifling of innovation.

To reemphasise the answer to RQ1, a new framework could address these issues by setting a clear focus and priority for safeguarding human rights, as suggested in the above section (6.2.1). The introduction of a focused framework on AI could offer clarity in these areas, and combine, as suggested, several ideas that exist within the literature to address ethical standards and accountability mechanisms. A new framework should take inspiration from the GDPR and be formed based on principles (see 6.2.3), which can be built upon to prioritise the importance of safeguarding human rights, such as the suggested FRIA (see 6.2.4). To ensure the strongest protection of human rights, future regulation should be outcome-focused based on the context of its use, rather than the general system or sector use itself. Effective regulation also needs effective redress mechanisms, and those suggested combine the ideas of placing duties of care, elements of strict liability, licensing, and a compensation scheme (see 6.2.7).

6.3.4 – Research Question 2

RQ2: *To what extent does the GDPR fail to account for the unique challenges posed by AI systems, and how could reform of the DPA drive more ethical AI use?*

Within this thesis, the connection between AI and the GDPR has formed a thread of discussion throughout. For the purposes of clarity, a future framework must take into account and align with other legislation, particularly the DPA, given the focus on AI systems that have personal data contained in their dataset or process personal data to decision-make. Although data protection regulation has helped to form the basis of the proposals made regarding a future framework for AI, adjustments to data protection rules would be needed to aid in this alignment.

To address RQ2, the elements of the DPA that are suggested in need of reform are the data protection principles of ‘purpose limitation’ and ‘data minimisation’.[1450] Due to the safeguards proposed, an exception to these principles in circumstances where the suggested framework would apply would allow for the cohesion of AI development in a more ethical, and practical way. Due to the suggested FRIA (see 6.2.4), the standards would be ‘outcome’ focused, and although the purpose is not limited, it is assessed in line with proportionality standards, to ensure the ethical use of AI is promoted. An exception to the data minimisation principle would also be necessary, to ensure that AI can use large datasets to expand its interpretation of data – however again, the FRIA would ensure such collection and use of data is proportionate and necessary for the context of use.

To address the lack of clarity in Articles 13-15,[1451] the right to explanation must be guaranteed for AI decision-making in a future framework, to provide legal certainty and effective redress mechanisms. As suggested, **transparency, explainability,** and **interpretability** should provide the basis of future regulation, making explanations mandatory, where it is possible to do so. To clarify the standard of an explanation given, the thesis advocates for counterfactual explanations, so that parties can determine the factors used behind decision-making. For circumstances where transparency is technically infeasible, additional safeguards should be put in place to balance protecting society, whilst avoiding the unnecessary stifling of innovation. This would include mandatory compliance with the suggested standard of human intervention (see 4.3.3 and 6.2.5 for further detail), and mandatory licensing before being put on the market (see 6.2.7).

[1450] GDPR (n 1) Article 5(1)(b) and (c).
[1451] ibid Articles 13-15.

As suggested, Article 22 must be refined to allow cohesion with AI, and to add legal certainty. In Chapter Four, proposals were made based on the standard of human intervention that should be mandated within legislation (see 4.3.3 and 6.2.5), and would include corporations having Intervention Officers, and logging documentation of their intervention decisions on a database shared with an Oversight Authority. In terms of interpreting the 'similarly significant' standard in Article 22, the thesis advocates for an objective standard with the burden of proof placed on the developers to release themselves from the restrictions within the article. This documentation would be reviewed through the suggested certification scheme, to ensure safeguards were able to be put in place where subjective components should be considered.

6.3.5 – Research Question 3

RQ3: *What limitations exist in recent attempts to regulate AI made by the EU and UK, and in reflection, what solutions can be proposed to strengthen safeguards whilst encouraging ethical innovation?*

The EU is clearly leading the way in regulating AI, with the first worldwide to introduce a blanket regulation. Although setting a standard, the Act has fallen short in several areas, including its loopholes to the prohibited systems under Article 5,[1452] the difficulties that stem from the risk-based approach and categorising systems, and the blanket exceptions to the regulation itself.

For effective regulation that promotes ethical AI, obligations must exist for all uses, regardless of the risk level posed, due to the unintended consequences of decisions. Blanket exceptions should also be avoided, particularly when historically, these have been used for disproportionate reasons and are often abused. To safeguard society, a balance must be struck between upholding human rights standards whilst avoiding the unnecessary stifling of innovation (see 6.1.1).

[1452] Provisional Agreement for the AI Act (n 103) Article 5.

Through suggestions made based on data protection regulation (see 6.2.2), principles should provide the basis of compliance (see 6.2.3), and interlink with obligations throughout a future framework, to ensure the strongest safeguards possible. Thorough monitoring and auditing mechanisms should be put in place both before deployment (see 6.2.4), and post-deployment to the market (see 6.2.5), and obligations should apply both to the public and private sector. It is important to note that the suggestions provided will not fill every gap that exists, however, can provide the basis for blanket regulation that can be built on in the future for specific sectors and uses of AI.

For regulation to work as intended, enforceability is paramount. The establishment of an AI authority is necessary to address the workload that will inevitably come with regulatory administration, and a new authority would not only create new jobs and opportunities, but would reduce the pressure and burden on other regulators, such as the ICO and Ofcom. The thesis advocates for an AI Human Oversight Board, which could assist in several areas (see 6.2.6), one of these being to aid with redress mechanisms. To uphold human rights standards, and to ensure victims of unpredictable or incorrect decisions can be compensated, measures for accountability must be set out clearly within a future framework (see 6.2.7).

It is only responsible to consider funding and education in line with such a regulation, to ensure that society and the industry can cope and adhere to the obligations set. As suggested, both Parliament and the Government can aid in ensuring educational and financial opportunities are made available to society (see 6.2.8), and to reap the benefits of the digital era; this should be made a priority. It is also important to consider the economic benefits to society through such regulations being introduced, such as new job opportunities relating to AI governance, oversight, auditing, and insurance. These opportunities, in line with education being promoted and encouraged through government investment, could help to counteract the ethical concern of job displacement caused by AI.

6.4 Contribution to Knowledge

The recommendations proposed above provide the foundation and purpose of this thesis, to propose up-to-date, justified, and well-researched suggestions that not only have the intention of safeguarding and protecting human rights, but also being practical and realistic in their application.

This thesis also stands out through its examination of the recent AI Act,[1453] only very recently published, making this project one of the first to examine the final published text in depth. Not only does this present new areas of examination, but it also raises discussion for improvements going forward in review of the changes made during the trilogue debates of the regulation. This research has developed in light of legal developments across the board, with the inclusion of the EU's establishment of the AI Office,[1454] and integration of domestic developments such as the Data Protection and Digital Information Bill (No2)[1455] and the Artificial Intelligence (Regulation) Bill.[1456]

England and Wales is in the very early stages of AI regulation, and although past statements have shown reluctancy in creating a new framework, it is clear that one is needed, a view somewhat supported by the Government in their White Paper,[1457] and in the House of Lords, through the introduction of an Artificial Intelligence (Regulation) Bill.[1458] For this reason, the thesis' contents act as a significant feature of what should be considered by England and Wales policymakers and regulators in the near future, particularly in terms of the reflections made on the lessons that can be learned from the AI Act.

Although Europe-wide legislation will soon be in effect, this thesis remains relevant to those yet to regulate (such as England and Wales), and the Europe-sphere itself to see where the final draft may fall short in expectations, particularly with a focus on the protection and safeguarding of human rights. The literature review provides a vast exploration and examination of the issues surrounding AI regulation, collating the bounds of research together into one single resource- providing an invaluable source for other researchers in the field. This thesis also considers global responses to AI regulation, providing a basis for other researchers and academics to learn

[1453] Provisional Agreement for the AI Act (n 103).
[1454] European Commission (n 1205).
[1455] Data Protection and Digital Information Bill (n 925).
[1456] Artificial Intelligence (Regulation) Bill (n 101).
[1457] Department for Science, Innovation and Technology (n 99).
[1458] Artificial Intelligence (Regulation) Bill (n 101).

about the similarities and differences in regulation, and to note the emerging themes, such as encouraging trustworthy AI.

It is also important to note that the AI Liability Directive has not yet been approved,[1459] and is currently in much earlier stages of the legislative process in comparison to the now finalised AI Act. In reference to this, the liability discussion contained within Chapters Two and Three, which this thesis has identified as a key factor to consider for future regulation, is still relevant for regulators across the board, and key considerations can be taken and understood in terms of the important of liability regimes being clarified.

The contents of the thesis could also be used as a reflection point in the future where sectorial legislation may need to fill the gaps where general regulation falls short, and can act as a basis of key themes of considerations, emphasising on the need to protect human rights. This research is also an invaluable and significant source to build upon in the future, by which the foundations have been set, allowing opportunities to research the points noted within the thesis further, in addition to venturing out to wider areas. Examples of such are explored in the below section (see 6.6).

Due to the research crossing multiple disciplines, the discussion bridges the gap between the legal, ethical, and technological sectors, providing an in-depth review of the past and current attempts at regulating AI. The argument within the thesis, to use inspiration from the GDPR as a basis for a future framework, has not been utilised through the EU's attempts to regulate AI, providing an opportunity for other regulators, such as England and Wales, to seek the advantages of the approach, one which is supported and justified throughout this research. For this reason, the significance and impact of this research should not be underestimated, particularly in light of the policy and practical implications available going forward.

6.5 Research Limitations

The discussion included within this thesis should be read in light of some limitations of the research. This section intends to highlight such limitations, with consideration

[1459] Proposed AI Liability Directive (n 104).

of the methodology and sources available, the topic area, the legal environment, and the novelty of the field.

The methodology chosen to underpin this research centres on qualitative secondary sources, combining a thematic analysis, doctrinal methodology and elements of macro and micro comparison, centred around a techno-legal-ethics approach. The reasoning behind this approach was based on the wealth of information already available in the public domain, including literature, legal proposals, debates, and decided case-law. In addition to this, due to public awareness and understanding of AI, particularly in the earlier stages of this project, primary research seeking views from laymen or those outside of the field would not have been useful to support the argument in the depth needed. Primary research to seek views from lawmakers and policymakers could have benefited the discussion, and could have brought in views which have not yet been considered in the literature or published materials. However, gaining access to such individuals would have been difficult, and due to the efforts in policymaking in this area, the consensus of views can be seen with transparency through the online domain. The combination of secondary source methodologies and approaches used also allowed for a thorough examination on the themes that ran throughout this thesis, and allowed for a holistic approach to consider the complexities that featured through taking an interdisciplinary, and well-rounded perspective.

The scope of this project has focused on fundamental human rights protection, with consideration of liability and general data protection aspects, but there are several other areas worthy of research that fall outside the scope of this project. For example, the conflict with Intellectual Property (IP) law, particularly concerning AI authorship,[1460] and the introduction of criminal provisions for copyright.[1461] Although in a very minor way, the EU's AI Act has picked up on the growing argument to clarify the rules on copyright, whereby in Recital 60f,[1462] it is stated that developers of general-purpose AI models are under an obligation to produce a summary of the

[1460] Case-145/10 *Painer v Standard VerlagsGmbH & Others* (2011) 62010CJ0145; P. Bernt Hugenholtz and João Pedro Quintais, 'Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?' [2021] 52 *IIC* 1190.

[1461] Felipe Romero Moreno and James GH Griffin, 'The UK's criminal copyright proposals in an era of technological precision' [2016] 7(2) *European Journal of Law and Technology* 1.

[1462] Provisional Agreement for the AI Act (n 103) Recital 60f content used for the training of the model, to align and respect copyright law.[1463] The AI Act also advises providers of general-purpose AI models to put in place a policy on copyright and related rights, to ensure copyright laws and obligations are followed.[1464]

Due to the thesis focusing on the fundamental rights aspect mainly considered within the core Convention rights,[1465] delving into this discussion of AI authorship and copyright alignment has fallen outside the scope of discussion, although may pave the way for future avenues of research to build on the points raised. The main rationale for this exclusion was the need to maintain a manageable and focused scope, allowing for more in-depth analysis of the issues where AI intersects with fundamental rights, liability and broader regulatory issues. Copyright law, whilst relevant to AI in specific contexts, primarily concerns IP rights associated with the creation and distribution of works by AI, and although could be connected to human rights, it more so presents a specialised area that warrants its own detailed exploration, separate to the ethical concerns and regulatory challenges addressed in this thesis.

Copyright and IP law only give one example of a topic area worthy of discussion falling outside of the scope of this project. Due to the wealth of information circulating relating to technological growth, the societal impact, and the intersectional issues between law and technology, this research equates to tackling a small part of a much bigger and widespread problem, and although this topic has been looked at in-depth, there are many other areas worth examining in the future.

This limitation also works in an opposing way; the topic area chosen as the focus covers the umbrella of AI when proposing framework recommendations. When beginning the research journey, public awareness towards the concept of AI was limited, and the technology did not feature as a part of public knowledge in the same way it does today. The aspect and growth of AI have led to differing categories that could be the basis of a thesis itself- such as regulating generative AI as opposed to all AI systems. Although the thesis does not explicitly focus on generative AI, the

[1463] ibid Recital 60f.
[1464] ibid Recital 60i and Article 52(C)(1)(c).
[1465] Human Rights Act 1998, Articles 2-14.

proposed regulatory framework is designed to be adaptable and inclusive of emerging AI technologies, including generative AI. The analysis given centres on core regulatory challenges such as ethical AI development, human rights concerns, and liability issues, challenges of which are equally relevant to generative AI. The interdisciplinary approach and recommendations within the thesis intend to accommodate the complexities of different AI systems, including generative AI. However, due to the nature of generative AI, it would be sensible to state that future obligations would be needed to capture the full elements specifically relevant to these types of systems.

To defend this, before regulating a niche area of AI, foundational legislation must first be in place and built upon; this is a process by which the law usually follows. Take for example human rights; the grounding legislation of the Human Rights Act provides a basis of protection,[1466] which has led to subsequent legislation, such as the Equality Act,[1467] to focus on an area within that foundational legislation, providing more context and in-depth rules, and filling the gaps within the foundational regulation.

Regulation must start somewhere, and it is most effective when building the ground rules and foundations for the future, before more niche areas are established. Due to the lack of widespread regulation currently in England and Wales, this thesis provides the basis for the current discussion regarding regulation, and avoids jumping too far into the future where such foundational regulation may make proposals for these future niche areas insignificant and meaningless.

The fast-paced legal environment whilst being a strength, also presented difficulties and limitations in writing this research. In principle, this research has taken place within a similar timespan of the EC's creation of the AI Act, and whilst this has provided a good basis for discussion and analysis, the fast-paced updates, proposals, and reports in the area have forced a continuous renewal of understanding and research. Due to the modern and recent changes in this area, it has been important to ensure that the research does not lose its longevity, and to ensure that the ideas within this thesis are not quickly outdated. To combat this, the

[1466] ibid.
[1467] Equality Act 2010.

research scope and discussions have had to change and adapt during the journey of being created, allowing a more in-depth and well-thought-out consideration of the key issues. This has also forced the research to be forward-thinking, and instead of settling on recommendations that may closely align with the EU's AI Act, the thesis has adapted to criticise the regulation, showing what lessons can be learnt for England and Wales during the EU's regulatory journey.

There are also limitations in the form of the absence of being able to examine the impact of the AI Act, due to the timing of completion of this research. The final draft of the AI Act was approved in early December of 2023, with the Act intended to be published in early 2024, aligning with the completion of this research. The research, therefore, has had to rely on speculations and agreements behind closed doors when considering the most recent developments to the AI Act, which can only be confirmed once official publication is made. Although it is beneficial that this research could take into consideration the final draft and decisions made by the EU in regulating AI, the short-term and long-term impact of the regulation is left uncalculated. Although EU law usually consists of a grace period before enforcement of regulation is put in place, there are expectations that the ban of systems will take place within six months of official publication. This means that, in the next six months following publication, the extent of the workings within Article 5 (of unacceptable risk systems)[1468] will be put into practice, and it would be interesting to measure how well the industry and technology developers cope with the restrictions put in place, particularly given scenarios where systems to be banned are currently on the market and in use.

Within the next two years, the full Act will be enforced. At this point, the enforcement mechanisms within the Act could be evaluated in practice, with an assessment of how the workings of the multiple bodies established (including the AI Office and EUAIB),[1469] and how well they collaborate to enforce and provide guidance on the regulation itself. After this time, the research could have extended to considering EU case-law under the AI Act, with an examination of how the courts deal with the Act, and any issues that may emerge through real-world scenarios and case studies.

[1468] Provisional Agreement for the AI Act (n 103) Article 5.

[1469] European Commission (n 1205); Provisional Agreement for the AI Act (n 103) Article 56.

Although the scope of this project could not capture these elements, these could provide a basis for future research in terms of extending the ideas introduced within the thesis, which reveals the longevity of this initial discussion and how it can be used as a foundation in the future.

The novelty of the field should also be noted; although the field of AI is rife in its research, particularly with the recent approval of the EU's AI Act,[1470] it should be highlighted how new and undeveloped AI technology is, and how the growth of AI is very early in its lifespan. Due to this, there may be areas of AI that are not quite yet understood in the same ways they will be in the future, and technological capabilities may increase to the point where, for example, black box systems presenting issues with explainability provisions are a problem of the past. Future developments may also cause complexities for the current discussion of regulation, which has been seen by the growth of generative AI, originally not fitting in appropriately to the early proposals of the AI Act.[1471] However, returning to the earlier point, to combat the pacing problem, regulation must start somewhere, and it is important to create the building blocks which can lead to regulatory developments in the future.

With consideration of these limitations, it is important to use these points identified as opportunities to think forward to areas in need of further exploration, and other avenues for future research, to ensure the longevity of the research and selected topic area, and to make an ongoing contribution to knowledge in the field and wider topic area. To take advantage of such opportunities, the next section of this conclusion delves into the future considerations of the field, assessing where the field of AI may develop, and proposing other avenues for further research.

6.6 Looking to the Future

To take advantage of such opportunities that arise from the limitations to the research identified in the above section, it is imperative, particularly given the nature of the topic of AI, to look forward to the future. This section reflects on the unanswered questions that remain following the research that has been carried out,

[1470] Provisional Agreement for the AI Act (n 103).
[1471] European Commission (n 117).

in addition to a summary of areas that need further exploration in light of the findings backed by this thesis.

For research such as this, a secondary source methodology can be advocated for. Although considering the limitations noted above, it is important to point out the advantages of such a methodology, particularly given the use of AI to increase and ease research, technological innovation, and the public awareness heightened towards developments in the field. As already noted, the common views of policymakers, corporations and academics are clear within the literature and parliamentary and institutional debates, so the information needed to collate and analyse is readily available in the public domain, particularly with the transparency given through the rule of law of the law-making process, taking away the need for primary research. Using a secondary source methodology also allows for a global perspective to be sought, comparing views, practices and policies generated around the world, to provide a more in-depth analysis of the domestic visions towards regulation.

Looking forward, future research into avenues that stem from this thesis could benefit and take advantage of the same methodology, providing a clear starting point for further study. In terms of considering the future of AI, it is interesting to reiterate the point mentioned earlier in this conclusion regarding how early the present world and society are in AI's lifespan, which gives room for consideration in terms of the extent to which AI technology will develop to in the future. Currently, there is the basis of an argument that Artificial General Intelligence (important to note that this refers to AGI rather than general-purpose AI) is close to being reached with the influx of large language models and generative AI. This thesis supports the view that although close, the Turing Test has not yet been met in these systems being indistinguishable from human counterparts,[1472] somewhat arguably, due to the models being trained to be that way. Take for example, a classic response from a large language model, or even a more basic system such as a voice assistant, who clearly reiterates the point that they are a machine, and not a human, so cannot carry out certain tasks or feel certain emotions.

[1472] Daniel Jannai, Amos Meron, Barank Lenz, Yoav Levine and Yoav Shoham, 'Human or Not? A Gamified Approach to the Turing Test' (2023) <https://arxiv.org/pdf/2305.20010.pdf> accessed 30th July 2023.

However, the likelihood of AI having the ability to break the barrier of AGI is not now absurd, or unlikely, which in the earlier stages and processes of this research, would not have necessarily been agreed with. With this in mind, the area of super-intelligence must be touched upon, and although the opinion behind the research towards the ability to create Artificial Super Intelligence (ASI) has been consistent throughout this process, and arguably will remain the same in the short-term future; a century from now, this might not so be the case. Particularly given the growth and technological innovation that has already taken place, it is believed that no limitations of what can be achieved should be put in place. It is important to consider here the distinguishment between what could be expected in the future, and the typical dystopian AI seen in movies and entertainment, which takes the discussion back to the importance of regulation being correct in these early stages. If the correct foundations are put in place now, it is hoped that in the future, regardless of the level of intelligence AI may achieve, the respect for human rights and the safeguarding of society will remain the same.

It is also important to consider this viewpoint from a global perspective, where human rights standards fluctuate, and government power and censorship in some nations are central to the workings of the institutions and the technological infrastructure. This leads to concern in terms of the products being created, and reaching the ‘wrong’ hands; a scenario which can be seen on a smaller scale today with the harm caused by technology being completely dependent on the scenario it is used in. For example, take the following examples as a basis for this discussion:

Scenario 1: Large language models being used to create and generate jokes for a family Christmas, whilst that same system is used by an isolated individual seeking medical advice in a jurisdiction where state-funded medical care is not given.

Scenario 2: A deepfake system being used to create filters on social media for entertainment, whilst also being used to create videos of political figures drawing up plans for conflict, or arbitrary policies.

These scenarios show a clear focus and emphasis that the use of systems is of significant importance, rather than the general system itself. This aligns with the overarching proposals above, that regulation must be outcome-focused, rather than based upon the basic and standard use of such systems. Looking at this from a broader perspective, a system which could be used proportionately in this jurisdiction, could be made readily available to other jurisdictions worldwide, that do not have the same level of human rights protection, nor clear rules or regulations governing such use of AI. This perhaps places a broader ethical obligation and responsibility for developers and deployers of systems, but it would be interesting to see, in terms of future research, whether there is the possibility for an establishment of rules on an international basis, similar to the UN documents, which although not having much legal influence, can influence countries politically and economically to ensure human right standards are sufficient worldwide.

In light of the above discussion, there are several different avenues of research as to which this thesis can be built upon including, for example, the following areas:

In-Depth Analysis of the Human Rights Impact on Specific Sectors

For example, particularly in reference to the human rights concern addressed earlier in this thesis, a notable extension of this research could lead to looking more specifically at the state use of AI, and the safeguards that are needed when AI is used to affect public institutions and infrastructures across the board. This would be a particularly fruitful discussion given the cases which have already appeared in the UK regarding state use of AI, whereby unfair or disproportionate decisions have been made. The *R(Bridges)*[1473] case is the most notable example in this aspect, but the Universal Credit algorithm used by the Government Department of Work and Pensions also gives a basis for such a discussion.[1474]

This could lead to delving deeper into more specific uses, particularly for example, law enforcement use of systems, or systems used for immigration purposes. This

[1473] *R(Bridges) v CC of South Wales Police* (n 49).

[1474] *Secretary of State for Work and Pensions v Danielle Johnson, Claire Woods, Erin Barrett, Katie Stewart* (n 743).

crosses over with the example case-law noted above (*R(Bridges)*[1475]) but could also link to law enforcement use examples from worldwide, where FRT has been integrated into US law enforcement body cameras, although due to widespread outrage, have since been removed.[1476] Particularly in the area of law enforcement, there has been long-standing concern regarding human rights abuses, and it is important to ensure that this is not worsened through the availability and use of AI. As an area of future research, it is hoped that the impact of the EU's AI Act would provide the basis of discussion, reflecting on how developers have adapted their systems to ensure they do not fall within the prohibited Articles, and how law enforcement react to such systems, assessing whether they are used appropriately under the Act, which could link to the case study included within Chapter Five on the use of FRT under the AI Act.

Another area that could factor into this section on future avenues of research could include consideration of the impact on elections; particularly given the rise of deepfake technology, targeted advertising, and generative AI. The area of generative AI and its impact on human rights has already formed the basis of further research from the interest gained through completing this thesis. The influx of 'fake news', or disinformation cannot be ignored, and the wide influx of disinformation usually links to political events, whether that be elections themselves,[1477] or global events such as the COVID-19 pandemic.[1478] With laymen being subject to masses of disinformation, manipulated content through deepfake technology, and the bias that exists behind generative AI, the impact on the freedom of thought could be the basis of consideration,[1479] in addition to the impact of the right to free and fair elections.[1480]

Another area that could be the basis of discussion is the use of AI in education. AI systems are impacting all sectors and industries, and the right to education is seemingly going to be affected by the influx of AI. The basis of future research could

[1475] *R(Bridges) v CC of South Wales Police* (n 49).
[1476] Axon AI and Policing Technology Ethics Board (n 162).
[1477] Nahema Marchal, Bence Kollanyi, Lisa-Maria Neudert, Hubert Au and Philip N. Howard, *Junk News & Information Sharing During the 2019 UK General Election* (Data Memo 2019.4, University of Oxford, 2019).
[1478] Aleksi Knuutila, Aliaksandr Herasimenko, Hubert Au, Jonathan Bright and Philip N. Howard 'A Dataset of COVID-Related Misinformation Videos and their Spread on Social Media' [2021] 7 Journal of Open Humanities Data 6
[1479] Human Rights Act 1998, Article 9 and 10.
[1480] ECHR (n 3) Article 3 of Protocol 1.

be examining whether AI would help to promote the right to education, or negatively impact upon it. This could be examined through consideration of already heavily criticised systems such as exam proctoring,[1481] listed within the EU's AI Act under Annex III as high risk,[1482] whilst also examining how AI has been used to tailor content to students, and whether any issues of bias have emerged through such usage, and how this impacts upon the right to an education.[1483] In this same sector, higher education institutions in particular have faced battles with students using large language models for answering assessments, which many universities now categorise as a form of academic misconduct.[1484] This could form the basis of research to examine how the education sector is responding to this issue, whether reluctance towards the use of the technology is the most appropriate response, and how, whether the use by students increases or decreases, the right to education will be impacted.

In-Depth Analysis of Specific AI Systems

The above discussion could then lead to research papers that assess specific systems used for specific means, such as generative AI in education, or facial recognition within the police force. In reflection of the AI Act, a particular area of interest would be the use of predictive policing systems, particularly given the loophole within the AI Act in the form of a partial ban under Article 5(da).[1485] Due to this only being a partial ban, it would be interesting to see how predictive policing systems remain in use, and whether the arguments stated earlier in terms of the human rights concerns remaining present, come to fruition. In the jurisdiction of England and Wales, this would be a particularly relevant discussion given the Metropolitan Police's use of the Gang-Matrix System,[1486] and the impact this has on local communities.

[1481] Simon Coghlan, Tim Miller and Jeannie Paterson, 'Good Proctor or "Big Brother"? Ethics of Online Exam Supervision Technologies [2021] 34 *Philosophy & Technology* 1581.

[1482] Provisional Agreement for the AI Act (n 103) Annex III (3).

[1483] ECHR (n 3) Article 2 of Protocol 1.

[1484] Geoffrey M. Currie, 'Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? [2023] 53(5) *Seminars in Nuclear Medicine* 719.

[1485] Provisional Agreement for the AI Act (n 103) Article 5(da).

[1486] Metropolitan Police, 'Gangs Violence Matrix' (Met. Police website) < https://www.met.police.uk/police-forces/metropolitan-police/areas/about-us/about-the-met/gangs-violence-matrix/> accessed 20th August 2023.

Another area of interest would be to assess the use of generative AI, particularly large language models, for particular uses. This would offer a great range of examination, as the use of such systems is determined by the end-user, rather than the developer or hosts of such a system. An area of curiosity would be where end-users use large-language models to gain advice; whether this would be legal advice, medical advice, or linking to the above discussion, advice on who to vote for in elections. This would widen the potential methodologies available for research, as systems such as ChatGPT are readily available, and primary research through using and testing responses from systems could be a basis for discussion. Particularly looking at the area of medical advice, this could invite discussions on the intersection of the Medical Devices Regulation,[1487] identifying any conflicts that would arise.

In-Depth Analysis of Specific Legal Issues

Away from a focus on the technology itself, other emerging technological areas that intersect with the law could be focused on for future research. An area of interest that has captured attention is the cross-section of healthcare AI and medical negligence, which has already led to scholarly outputs through the interest raised during the research process of this thesis. The basis of such a discussion could build upon the ideas within this thesis in terms of regulation, and how laws may need to be adapted to accommodate for a post-AI world. The examination of precedents set by the courts would also allow for a basis of discussion here; for example, in reference to medical negligence, the longstanding Bolam and Bolitho test would be unusable with the integration of AI,[1488] particularly if used in decision-making circumstances.

With the widespread growth and developments in the field of AI, future potential areas of review are limitless, with the above giving only a few examples of areas that have caught awareness and attention through the completion of this thesis. This thesis intends to act as a building block to further research, which can offer a wide

[1487] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [2017] OJ L 117/1.

[1488] *Bolam v Friern Hospital Management Committee* (n 872); *Bolitho v City and Hackney Health Authority* (n 873).

range of reflections in the future, where other topics and areas are being examined and reviewed.

6.7 Concluding Points

This project has been methodologically innovative in the ways it has bridged the gap between the legal, ethical, and technological fields, through the consideration of past, recent and current attempts at regulating AI. This research has provided an invaluable source for the field of AI and law, and has contributed to the research of the emerging digital world and society in terms of the focus on respecting and upholding human rights standards. The research has provided an in-depth analysis of the complexities of AI technology, and the conflicts caused by current regulatory regimes, arguing that the foundations within the GDPR should give a basis for a future framework for AI within England and Wales. This novel perspective can aid in advancing the value of effective regulation in the field, and provides a basis to build upon for future sectorial law.

The recommendations that have been proposed within this thesis are significant, providing an insight into the most recent regulatory developments, to ensure such solutions are up-to-date, justified, well-researched, realistic, and practical. These solutions act as an alternative approach to the EU's AI Act, and aim to improve on the areas where the EU's attempts have fallen short.

The thesis is unique in the way it opens up a platform and basis for future study, ensuring that the predominant ideas that have been suggested continue to live on in future papers, projects, and research. The thesis conducted has a real-world relevance, not only for the basis of future research, but also for shaping policy and decisions on regulating the use of AI. Future regulation must be clear and concise, whilst practical and flexible, to strike the balance between promoting ethical AI, whilst also ensuring that human rights standards are at the forefront of decision and policymaking, offering society the opportunity to live in an AI world that benefits them.

Table of Case-Law:

England and Wales:

A and Others v National Blood Authority and another [2001] 3 All ER 289

Anns v Merton London Borough Council [1978] AC 728

Associated Provincial Picture Houses Ltd v *Wednesbury* Corporation (1948) 1 KB 223

Barnett v Chelsea & Kensington Hospital [1969] 1 QB 428

Bolam v Friern Hospital Management Committee [1957] 1 WLR 582

Bolitho v City and Hackney Health Authority [1996] 4 All ER 771

Caparo Industries plc v Dickman [1990] 2 AC 605

Campbell v Mirror Group Newspapers Ltd [2004] UKHL 22 [2004] 2 WLR 1232

Carlill v Carbolic Smoke Ball Company [1892] EWCA Civ 1

Chester v Afshar [2004] UKHL 41 [2004] 10 WLUK 38

Cox v Ministry of Justice [2016] UKSC 10 [2016] 3 WLUK 91

Donoghue v Stevenson [1932] AC 562

Groom v Croker [1939] 1 KB 194

Hinz v Berry [1970] 2 QB 40

Honeywill and Stein Ltd v Larkin Brothers Ltd [1934] 1 KB 191

Lister v Hesley Hall Ltd [2001] UKHL 22 [2001] 5 WLUK 105

Mcquire v Western Morning News [1903] 2 KB 100

Mohamud v WM Morrison Supermarkets Plc [2016] UKSC 11 [2016] 3 WLUK 90

Montgomery v Lanarkshire Health Board [2015] UKSC 11 [2015] 3 WLUK 306

Page v Smith [1996] 1 AC 155

Pippin v Sheppard (1822) 147 ER 512

R v Dalloway [1847] 2 Cox 273

R v White [1910] 2 KB 124

R(Bridges) v Chief Constable of South Wales [2019] EWCH 2341 (Admin) [2020] 1 WLR 672

R(Bridges) v Chief Constable of South Wales [2020] EWCA Civ 1058 [2020] 1 WLR 5037

Robinson v Chief Constable of West Yorkshire Police [2018] UKSC 4 [2018] WLR 595

Roe v Minister of Health [1954] 2 WLR 915

Salomon v Salomon & Company Ltd [1897] AC 2

Secretary of State for Work and Pensions v Danielle Johnson, Claire Woods, Erin Barrett, Katie Stewart [2020] EWCA Civ 778 [2020] 6 WLUK 270

The Wagon Mound no 1 [1961] AC 388

Yewens v Noakes (1881) 6 QBD 530

ECtHR:

Barbulescu v Romania. App no 61496/08 (ECtHR, 5 September 2017) ECHR 754

C.R. v the UK. App No 20190/92 (ECtHR, 22 November 1995) A/3550-C

Catt v the UK. App no 43514/15 (ECtHR, 24 April 2019)

Klass and others v Germany. App no 5029/71 (ECtHR, 6 September 1978) 2 EHRR 214

Leander v Sweden. App no 9248/81 (ECtHR, 26 March 1987) A/116

Rotaru v Romania. App no 28341/95 (ECtHR, 4th May 2000)

S. and Marper v the UK. App no 30562/04 and 30566/04 (ECtHR, 4 December 2008) 48 E.H.R.R 50

Satakunnan Markkinaporssi Oy v Finland. App No 931/13 (ECtHR, 27 June 2017)

Szabó and Vissy v. Hungary. App no 37138/14 (ECtHR, 12 January 2016)

European Union:

Case 215/88 Casa Fleischhandels-GmbH v Bundesanstalt für landwirtschaftliche Marktordnung [1989] ECR-2789.

Case C-184/20 OT v Vyriausioji tarnybinės etikos komisija [2022] 62020CJ0184.

Case C-228/21 Request for a preliminary ruling from the Corte suprema di cassazione (Italy) lodged on 8 April 2021 — Ministero dell'Interno, Dipartimento per le Libertà civili e l'Immigrazione — Unità Dublino v CZA (2021) OJ C 217/31.

Case C-28/08 Commission v Bavarian Lager [2010] ECR 2010 I-06055.

Case C-311/18 Data Protection Commissioner v Facebook Ireland Limited and Maximilliam Schrems [2020] OJ C 249.

Case C-401/19 Poland v Parliament and Council [2022] OJ C 270.

Case C-434/16 Nowak v Data Protection Commissioner [2017] 62016CJ0434.

Case C-817/19 Ligue des droits humains v Conseil des ministers [2020] OJ C 36/16.

Case-145/10 *Painer v Standard VerlagsGmbH & Others* (2011) 62010CJ0145.

Joined Cases C-141/12 and C-372/12 Y.S v Minister voor Immigratie, Integratie en Asiel, and Minister voor Immigratie, Integratie en Asiel v M. and S. [2014] 62012CJ0141.

Joined Cases C-203/15 and C-698/15 Tele2 Sverige AB v Post-och telestyrelsenk [2016] All ER (D) 107 and Secretary of State for the Home Department v Tom Watson [2016] All ER (D) 107.

Europe:

Algorithm Transparency Case (Garante per la Protezione dei Dati PersonaliAssociazione Mevaluate Onlus) (2021) Case 14381/2021 (Italy)

Dutch Legal Committee for Human Rights v State of the Netherlands [2020] C/09/550982 HA ZA 18-388 (Netherlands)

eKasa System Case (Ústavného súdu Slovenskej republiky) (2021) Case 492/2021 Z. z. (Slovakia)

Ola v Ola Drivers [2021] C/13/689705/HA RK 20-258 (Netherlands)

Uber v Uber Drivers (Deactivation Case) [2021] C/13/692003/HA RK 20-302 (Netherlands)

Uber v Uber Drivers (Transparency Case) [2021] C/13/687315/HA RK 20-207 (Netherlands)

United States:

Citizens United v Federal Election Commission. 558 US 310 (2010)

Loomis v Wisconsin. 137 S.Ct. 2290 (2017)

Santa Clara County v Southern Pacific Railroad. 118 U.S. 394 (1886)

Table of Legislation, Bills and Treaties:

Bills:

Artificial Intelligence (Regulation) Bill, HL Bill (2023-24)

Data Protection and Digital Information Bill (No.2) HC Bill (2022-23, 2023-24)

England and Wales:

Data Protection Act 2018

Equality Act 2010

Human Rights Act 1998

National Security and Investment Act 2021 (Notifiable Acquisition) (Specification of Qualifying Entities) Regulations 2021

Online Safety Act 2023

Road Traffic Act 1988

European Union:

Charter of Fundamental Rights of the European Union [2000] OJ C364/1.

Consolidated version of the Treaty establishing the European Community [2002] OJ C 325.

Council Directive (EC) 2009/136 concerning the processing of personal data and the protection of privacy in the electronic communications sector OJ L 337.

Council Directive (EC) 85/374 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products (Product Liability Directive) [1985] OJ L 210.

Council Directive (EC) 95/46 of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive) [1995] OJ L 281.

Council Directive (EC) relating to insurance against civil liability in respect of the use of motor vehicles, and the enforcement of the obligation to insure against such liability (2009/103/EC) OJ L 263.

Regulation (EU) 19/1150 on promoting fairness and transparency for business users of online intermediation services [2019] OJ L 186.

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

Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and

Regulation (EU) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [2017] OJ L 117/1.

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) [2022] OJ L 265.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277.

Human Rights:

Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14 1950 (ECHR).

International Covenant on Economic, Social and Cultural Rights (ICESCR) 1966.

Universal Declaration of Human Rights (adopted 10 December 1948 UNGA Res 217 A(III) (UDHR).

Non-Domestic Jurisdictions:

Accident Compensation Act 2001 (New Zealand).

Algorithmic Accountability Act of 2022 (H.R.6580), 117th Congress (2021-2022) (US).

Amendment of the Act on Protection of Personal Information 2020 (Act No. 57 of 2003 as amended in 2015) (APPI) (Japan’s Data Protection Act).

American Data Privacy and Protection Act (H.R. 8152), 117th Congress (2021-2022) (US).

An act to add Chapter 5.9 (commencing with Section 11549.80) to Part 1 of Division 3 of Title 2 of, the Government Code, relating to state government (Senate Bill No. 313) 2023 (California).

An Act to amend Sections 12930 and 14203 of the Government Code,and to amend Section 156 of 90.5 of, and to add Part 5.6 (commencing with Section 1520) to Division 2 of, the Labor Code, relating to employment (Assembly Bill No. 1651) 2022 (California).

Artificial Intelligence Bill (translated) (Federal Senate Bill, No. 21 of 2020) (Brazil).

Consumer Privacy Protection Act 2022 Bill C-27 (44th Parliament, 1st Session, 2021-2022) (Canada).

Digital Charter Implementation Act 2022 C-27 (44th Parliament, 1st Session, 2021-2022) (Canada).

Directive on Automated Decision-Making 2019 (Canada).

Health Equity and Accountability Act of 2022 (H.R. 7585), 117th Congress (2021-2022) (US).

Lei Geral de Proteção de Dados (Federal Law no. 12,709/2018) (Brazil's Data Protection Act).

Loi pour une République numérique (Digital Republic Act, 2016-1321) (France).

National Defense Authorization Act for Fiscal Year 2023 (S.4543), 117th Congress (2021-2022) (US).

Oakland Municipal Code, Ordinance adding Chapter 9.64 Establishing Rules for the City's Acquisition and Use of Surveillance Equipment 2018 (Oakland, US).

Bibliography:

Academic Blogs:

Abbass H. 'An AI professor explains: three concerns about granting citizenship to robot Sophia' (The Conversation, 30 October 2017) <https://theconversation.com/an-ai-professor-explains-three-concerns-about-granting-citizenship-to-robot-sophia-86479> accessed 13th February 2019.

Angwin J, Larson J, Mattu S and Kirchner L, 'Machine Bias' (ProRepublica, 23rd May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 11th January 2021.

Bacciarelli A, 'Ethical AI principles won't solve a human rights crisis' (Amnesty International, 21st June 2019) <https://www.amnesty.org/en/latest/research/2019/06/ethical-ai-principles-wont-solve-a-human-rights-crisis/> accessed 14th November 2020.

Barkved K, 'How to know if your machine learning model has good performance' (Obviously AI, 9th March 2022) <https://www.obviously.ai/post/machine-learning-model-performance> accessed 20th May 2022.

Barlow Keener E, 'Facial Recognition: A New Trend in State Regulation' (Womble Bond Dickinson Blog, 29th April 2022) <https://www.womblebonddickinson.com/us/insights/alerts/facial-recognition-new-trend-state-regulation> accessed 2nd October 2023.

Baweja S, and Singh S, 'Beginning of Artificial Intelligence, End of Human Rights' (LSE Blog, 16th June 2020) <https://blogs.lse.ac.uk/humanrights/2020/07/16/beginning-of-artificial-intelligence-end-of-human-rights/ > accessed 15th November 2020.

Bellon T, 'Liability and legal questions follow Uber autonomous car fatal accident' (Insurance Journal, 20 March 2018) <https://www.insurancejournal.com/news/national/2018/03/20/483981.htm> accessed 24th March 2019.

Bergan B, 'Germany Drafts World's First Ethical Guidelines for Self-Driving Cars' (Futurism Blog, 25 August 2017) < https://futurism.com/germany-drafts-worlds-first-ethical-guidelines-for-self-driving-cars> accessed 22nd June 2020.

Bert A, 'The AI Transparency Paradox' (Harvard Business Review Blog, 13th December 2019) <https://hbr.org/2019/12/the-ai-transparency-paradox> accessed 4th March 2020.

Berthelemy C, 'UN Special Rapporteur Analyses AI's Impact on Human Rights' (European Digital Rights Blog, 7th November 2018) <https://edri.org/un-special-rapporteur-report-artificial-intelligence-impact-human-rights/> accessed 4th March 2020.

Bertuzzi L, 'Czech Presidency sets out path for AI Act discussions' (Euractiv Blog, 22nd June 2022, updated 28th June 2022) <https://www.euractiv.com/section/digital/news/czech-presidency-sets-out-path-for-ai-act-discussions/> accessed 18th December 2022.

Bertuzzi L, 'Leading MEPs raise the curtain on draft AI rules' (Euractiv Blog, 11 April 2022, Updated 13th April 2022) <https://www.euractiv.com/section/digital/news/leading-meps-raise-the-curtain-on-draft-ai-rules/> accessed 2nd January 2023.

Bertuzzi L, 'Making the AI Act work for SMEs: The EU tries to square the circle' (Euractiv Blog, 25th November 2022) <https://www.euractiv.com/section/digital/news/making-the-ai-act-work-for-smes-the-eu-tries-to-square-the-circle/> accessed 22nd July 2023.

Bertuzzi L, 'The US unofficial position on upcoming EU Artificial Intelligence Rules' (Euractiv Blog, 24th October 2022, updated 26th October 2022) <https://www.euractiv.com/section/digital/news/the-us-unofficial-position-on-upcoming-eu-artificial-intelligence-rules/> accessed 8th January 2023.

Bickerstaff R, and Mohan A, 'A 'Light Touch' Regulatory Framework for AI – Transparency at the Heart of AI Regulation' (Digital Business Law Lexology Blog, ND) <https://www.lexology.com/library/detail.aspx?g=33697021-f272-4d51-ad0f-2e1d87e0857d> accessed 14th January 2021.

Big Brother Watch, 'Response to Court of Appeal Judgment in Dr Bridges' Challenge' (11th August 2020) <https://bigbrotherwatch.org.uk/2020/08/big-brother-watchs-response-to-court-of-appeal-judgment-in-dr-bridges-challenge-to-live-facial-recognition/> accessed 5th November 2020.

Binns R, and Gallo V, 'Human bias and discrimination in AI systems' (Wired Gov Blog, 26th June 2019) <https://www.wired-gov.net/wg/news.nsf/articles/Human+bias+and+discrimination+in+AI+systems+26062019152000?open> accessed 5th November 2020.

Browne G, 'The Fall of Babylon Is a Warning for AI Unicorns' (Wired Blog, 19th September 2023) <https://www.wired.co.uk/article/babylon-health-warning-ai-unicorns> accessed 4th January 2022.

Burt A, 'Is There a Right to Explanation for Machine Learning in the GDPR?' (Privacy Tech Blog, 1st June 2017) <https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/> accessed 22nd December 2019.

Campbell C, 'The EU's Draft AI Regulations: A Thoughtful But Imperfect Start' (Medium Blog, 13th August 2021) <https://catriona-campbell.medium.com/the-eus-draft-ai-regulations-a-thoughtful-but-imperfect-start-8c6489f1617> accessed 15th September 2023.

Cankett M and Liddy B, 'Risk management in the new era of AI regulation' (Deloitte Blog: Audit & Assurance, 12th July 2022)

<https://www2.deloitte.com/uk/en/blog/auditandassurance/2022/the-new-era-of-ai-regulation.html> accessed 14th September 2023.

Castelvecci D, 'AI Pioneer: "The dangers of abuse are very real" (Nature Research, 4th April 2019) <https://www.nature.com/articles/d41586-019-00505-2> accessed 6th February 2020.

Castro D and Chivot E, 'How the EU should revised its white paper before its published' (Center for Data Innovation, 1st February 2020) <https://datainnovation.org/2020/02/how-the-eu-should-revise-its-ai-white-paper-before-it-is-published/> accessed 4th December 2020.

Castro D and Chivot E, 'Want Europe to have the best AI? Reform the GDPR' (Privacy Perspectives, 23rd May 2019) <https://iapp.org/news/a/want-europe-to-have-the-best-ai-reform-the-gdpr/> accessed 12th May 2020.

Castro D and McLaughlin M, 'Ten Ways the Precautionary Principle Undermines Progress in Artificial Intelligence' (Information Technology and Innovation Foundation, 4th February 2019) <https://itif.org/publications/2019/02/04/ten-ways-precautionary-principle-undermines-progress-artificial-intelligence> accessed 16th November 2020.

Copeland E, 'Does the public sector really need a code of AI ethics? (Nesta Blogs-Government Innovation, 15th February 2019) <https://www.nesta.org.uk/blog/does-public-sector-really-need-code-ai-ethics/> accessed 15th March 2020.

Crumpler W, 'Europe's Strategy for AI Regulation' (CSIS Blog, 21st February 2020) <https://www.csis.org/blogs/strategic-technologies-blog/europes-strategy-ai-regulation> accessed 20th July 2023.

Dastin J, 'Amazon scraps secret AI recruiting tool that showed bias against women' (Reuters Blog, 11th October 2018) <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> accessed 4th November 2020.

DataEthics, 'GDPR Does Entail a Right to Explanation' (Dataethics Blog, 2nd December 2017) <https://dataethics.eu/study-gdpr-entail-right-explanation/> accessed 4th March 2020.

Daws R, 'UK Police are concerned AI will lead to bias and over-reliance on automation' (Artificial Intelligence News Blog, 17th September 2019) <https://artificialintelligence-news.com/2019/09/17/uk-police-concerned-ai-bias-automation/> accessed 30th January 2020.

Delcker J, 'Europe divided over robot 'personhood' (Politico Blog, 11th April 2018) < https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/> accessed 20th March 2019.

Ebert A, 'We Want Fair AI Algorithms – But How To Define Fairness?' (Mostly AI Blog, 6th May 2020) <https://mostly.ai/2020/05/06/we-want-fair-ai-algorithms-but-how-to-define-fairness/> accessed 19th January 2021.

Engler A, 'The EU AI Act will have global impact, but a limited Brussels Effect' (Brookings Blog, 8th June 2022) <https://www.brookings.edu/articles/the-eu-ai-act-will-have-global-impact-but-a-limited-brussels-effect/> accessed 5th August 2023.

Fai M, Bradley J and Kirker E, 'Lessons in 'Ethics by Design' from Britain's A Level algorithm' (Gilbert and Tobin, Lexology, 11th September 2020) <https://www.lexology.com/library/detail.aspx?g=af0fcf0c-8f56-4f8e-ab7b-c3ece59bfdf5> accessed 20th September 2023.

Fanucci F and Connolly C, 'What are the AI Act and the Council of Europe Convention' (Stop Killer Robots, 14th August 2023) <https://www.stopkillerrobots.org/news/what-are-the-ai-act-and-the-council-of-europe-convention/> accessed 1st September 2023.

Freeman R, Temple C, Bischofberger T, Dobson SJ and Roberts C, 'Product liability and safety in the EU: overview' (Practical Law, 1st August 2020) < https://uk.practicallaw.thomsonreuters.com/w-013-0379?transitionType=Default&contextData=(sc.Default)&firstPage=true> accessed 13th November 2020.

Garvie C, 'Garbage in, Garbage out' (Georgetown Law, Center on Privacy and Technology, 16th May 2019) <https://www.flawedfacedata.com/> accessed 17th November 2020.

Gill J, 'Where Does eDiscovery Fit in the Facial Recognition Conversation?' (JD Supra Blog, 5th August 2019) <https://www.jdsupra.com/legalnews/where-does-ediscovery-fit-in-the-facial-40933/> accessed 14th June 2020.

Gluyas L and Day S, 'Artificial Intelligence- who is liable when AI fails to perform?' (CMS Blog, Law and Tax, 2018) <https://cms.law/en/GBR/Publication/Artificial-Intelligence-Who-is-liable-when-AI-fails-to-perform> accessed 24th March 2019.

Grady P, 'EU's AI Act Resurrects Subliminal Messaging Panic' (Center for Data Innovation Blog, 21st October 2022) <https://datainnovation.org/2022/10/eus-ai-act-resurrects-subliminal-messaging-panic/> accessed 2nd September 2023.

Gulley A and Hilliard A, 'The Proposed Amendments to California's Employment Legislation Regarding Automated-Decision Systems' (Holistic AI, 18th June 2023) <https://www.holisticai.com/blog/california-employment-legislation-proposed-amendments> accessed 2nd November 2023.

Habersetzer N, 'Russian Activists Fights Use of Facial Recognition Technology' (Human Rights Watch, 18th October 2019) <https://www.hrw.org/news/2019/10/18/russian-activist-fights-use-facial-recognition-technology> accessed 17th November 2022.

Hadfield G, 'Rules for Robots: The Path to Effective AI Regulation' (MIT Digital Blog, 12th June 2019) <http://ide.mit.edu/news-blog/blog/rules-robots-path-effective-ai-regulation> accessed 21st February 2020.

Hardesty L, 'Study finds gender and skin-type bias in commercial artificial-intelligence systems' (MIT News, 11th February 2018) <https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212> accessed 10th November 2020.

Hare S, 'It's time for AI ethics to grow up' (Wired Blog, 8th January 2020) <https://www.wired.co.uk/article/ai-ethics-law> accessed 4th November 2020.

Harkens A, 'Not just A-levels: unfair algorithms are being used to make all sorts of government decisions' (The Conversation, 3rd September 2020) <https://theconversation.com/not-just-a-levels-unfair-algorithms-are-being-used-to-make-all-sorts-of-government-decisions-145138> accessed 18th September 2023.

Hartwig B, 'The Impact of Artificial Intelligence on Human Rights' (Dataversity, 8th May 2020) <https://www.dataversity.net/the-impact-of-artificial-intelligence-on-human-rights/> accessed 16th November 2020.

Hidvegi F and Leufer D, 'European Union: more big words on AI, but where are the actions?' (Press Release, 26th June 2019) <https://www.accessnow.org/european-union-more-big-words-on-ai-but-where-are-the-actions/> accessed 19th November 2020.

Hidvegi F and Leufer D, 'Trust and excellence – the EU is missing the mark again on AI and human rights' (Access Now Blog, 11th June 2020) <https://www.accessnow.org/trust-and-excellence-the-eu-is-missing-the-mark-again-on-ai-and-human-rights/> accessed 29th November 2020.

Holland Michel A, 'Inside the messy ethics of making war with machines' (MIT Technology Review Blog, 16th August 2023) <https://www.technologyreview.com/2023/08/16/1077386/war-machines/> accessed 1st September 2023.

Hosanagar K, 'As machines become more intelligent, they also become unpredictable' (FoundingFuel Article, 2nd August 2019) <https://www.foundingfuel.com/article/as-machines-become-more-intelligent-they-also-become-unpredictable/> accessed 20th January 2021.

Human Rights Watch, 'UK: Automated Benefits System Failing People in Need' (Blog, 29th September 2020) <https://www.hrw.org/news/2020/09/29/uk-automated-benefits-system-failing-people-need> accessed 19th January 2021.

Johnson J, 'Interpretability vs Explainability: The Black Box of Machine Learning' (BMC Blog, 16th July 2020) <https://www.bmc.com/blogs/machine-learning-interpretability-vs-explainability/> accessed 15th January 2021.

Kelly A, 'The great algorithm fiasco' (BERA Blog, 21st May 2021) <https://www.bera.ac.uk/blog/the-great-algorithm-fiasco> accessed 20th September 2023.

Klein P, 'Rules for Robots: The Path to Effective AI Regulation' (MIT Digital Blog, 12th June 2019) <http://ide.mit.edu/news-blog/blog/rules-robots-path-effective-ai-regulation> accessed 21st February 2020.

Laranjeira de Pereira JR and Moraes TG, ‘Promoting irresponsible AI: lessons from a Brazilian bill’ (Heinrich-Böll-Stiftung- German political foundation blog, 14th February 2022) <https://eu.boell.org/en/2022/02/14/promoting-irresponsible-ai-lessons-brazilian-bill> accessed 5th November 2023.

Larson J, Mattu S, Kirchner L and Angwin J, ‘How we Analyzed the COMPAS Recidivism Algorithm’ (23rd May 2016) <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> accessed 12 January 2021.

Lee P, ‘Learning from Tay’s introduction’ (Microsoft Blog, 25th March 2016) <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/> accessed 8th November 2020.

Leprince-Ringuet D, ‘Even computer experts think ending human oversight of AI is a very bad idea’ (14th October 2021) <https://www.zdnet.com/article/even-computer-experts-think-ending-human-oversight-of-ai-is-a-very-bad-idea/> accessed 19th June 2022.

Leufer D and Jakubowska E ‘Attention EU regulators: we need more than AI ‘ethics’ to keep us safe’ (Access Now Blog, 21st October 2020) <https://www.accessnow.org/eu-regulations-ai-ethics/> accessed 29th November 2020.

MacCarthy M and Propp K ‘The EU’s White Paper on AI: A Thoughtful and Balanced Way Forward’ (LawFare Blog, 5th March 2020) <https://www.lawfareblog.com/eus-white-paper-ai-thoughtful-and-balanced-way-forward> accessed 3rd December 2020.

Manyika J, Silberg J and Presten B, ‘What Do We Do About the Biases in AI?’ (Harvard Business Review, 25th October 2019) <https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai> accessed 5th November 2020.

Marchant G, ‘”Soft Law” Governance of Artificial Intelligence’ (AI Pulse Blog, 25th January 2019) <https://escholarship.org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf?t=po1uh8> accessed 4th February 2020.

Marsden C and Nicholls R, ‘Interoperability: A solution to regulating AI and social media platforms’ (Tech Law for Everyone Blog, ND) <https://www.scl.org/articles/10662-interoperability-a-solution-to-regulating-ai-and-social-media-platforms> accessed 5th March 2020.

Mate P, ‘Branches of Artificial Intelligence’ (My Road to Artificial Intelligence, Medium Blog, 25th May 2020) <https://medium.com/myroadtoartificialintelligence/branches-of-artificial-intelligence-812b8e292cdb> accessed 19th August 2020.

McIlroy J, ‘Toyota boss says autonomous cars will accept crash liability’ (AutoExpress, 22nd October 2019) < https://www.autoexpress.co.uk/toyota/108205/toyota-boss-says-autonomous-cars-will-accept-crash-liability> accessed 13th November 2020.

McMullen M, 'How to Improve Accuracy of Machine Learning Model?' (Medium, 19th August 2019) <https://cogitotech.medium.com/how-to-improve-accuracy-of-machine-learning-model-5ee122727dc1> accessed 22nd September 2022.

Mehta A, 'What is Accuracy, Precision, Recall and F1 score? What is its significance in Machine Learning?' (Medium Blog, 17th September 2020) <https://medium.com/ai-in-plain-english/what-is-accuracy-precision-recall-and-f1-score-what-is-its-significance-in-machine-learning-77d262952287> accessed 22nd January 2021.

Menn J, 'Microsoft tuned down facial-recognition sales on human right concerns' (Reuters, 17th April 2019) <https://www.reuters.com/article/idUSKCN1RS2FX/#:~:text=Microsoft%20concluded%20it%20would%20lead,mostly%20white%20and%20male%20pictures> accessed 17th November 2020.

Meyer D, 'Europe thinks ethics is the key to winning the AI race. Not everyone is convinced' (Fortune Blog, 8th April 2019) <http://fortune.com/2019/04/08/eu-ai-ethics-principles/> accessed 10th April 2019.

Mijatović D, 'In the era of artificial intelligence: safeguarding human rights' (Open Democracy Blog, 3rd July 2018) <https://www.opendemocracy.net/en/digitaliberties/in-era-of-artificial-intelligence-safeguarding-human-rights/> accessed 15th November 2020.

Mundell I, 'Healthcare AI' (Imperial College London, October 2019) <https://www.imperial.ac.uk/enterprise/long-reads/healthcare-ai/> accessed 12th November 2020.

Narayanan A, 'A Human Rights Framework is Necessary to Govern Artificial Intelligence' (Human Rights Pulse, 8th June 2020) <https://www.humanrightspulse.com/mastercontentblog/a-human-rights-framework-is-necessary-to-govern-artificial-intelligence> accessed 14th November 2020.

Nelson M and Reinhold F, 'The DSA Proposal is a good start. Now policymakers must ensure that it has teeth' (Algorithm Watch, 16th December 2020) <https://algorithmwatch.org/en/dsa-response/> accessed 17th January 2021.

Oleksiuk A, 'How to train AI with GDPR limitations' (Intellias Intelligent Software Engineering Blog, 13th September 2019) <https://www.intellias.com/how-to-train-an-ai-with-gdpr-limitations/> accessed 20th August 2020.

Oxborough C and Cameron E, 'Explainable AI' (PWC, 2018) <https://www.pwc.co.uk/services/risk-assurance/insights/explainable-ai.html> accessed 3rd November 2020.

Powles J, 'New York City's Bold, Flawed Attempt to Make Algorithms Accountable' (The New Yorker, 21st December 2017) <https://www.newyorker.com/tech/annals-of-

technology/new-york-citys-bold-flawed-attempt-to-make-algorithms-accountable> accessed 17th November 2020.

Protalinksi E, 'ProBeat: Why Google is really calling for AI regulation' (Venture Beat Blog, 24th January 2020) <https://venturebeat.com/2020/01/24/probeat-why-google-is-really-calling-for-ai-regulation/> accessed 19th May 2020.

Reeve O, 'Celebrating Ada Lovelace Day: what Ada means to us' (Ada Lovelace Institute, 8th October 2019) <https://www.adalovelaceinstitute.org/blog/celebrating-ada-lovelace-day/> accessed 15th November 2020.

Richmond L, 'Artificial Intelligence: Who's to blame?' (Tech Law for Everyone, 8th August 2018) <https://www.scl.org/articles/10277-artificial-intelligence-who-s-to-blame> accessed 12th November 2020.

Ross C, 'IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show' (Stat Blog, 25th July 2018) <https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/> accessed 18th January 2021.

Saucedo A, 'The top three risks posed by AI, and how to safeguard against them' (IR Pro Portal Blog, 14th August 2020) <https://www.seldon.io/itproportal-the-top-three-risks-posed-by-ai-and-how-to-safeguard-against-them> accessed 5th November 2020.

Schwartz O, 'In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation' (IEEE Spectrum Blog, 25th November 2019, updated 4th January 2024) <https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation> accessed 8th November 2020.

Sharkey N, 'The impact of gender and race bias in AI' (Humanitarian Law & Policy Blog, 28th August 2018) <https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/> accessed 12th January 2021.

Shin T, 'Real-life Examples of Discriminating Artificial Intelligence' (Towards Data Science Blog, 4th June 2020) <https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070> accessed 12th January 2021.

Skelton SK, 'Europe's proposed AI regulation falls short on protecting rights' (Computer Weekly Blog, 14th June 2021) <https://www.computerweekly.com/feature/Europes-proposed-AI-regulation-falls-short-on-protecting-rights> accessed 12th September 2023.

Snow J, 'Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots' (American Civil Liberties Union Blog, 26th July 2018) <https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28> accessed 4th February 2020.

Spyridaki K, 'GDPR and AI: Friends, Foes or something in between?' (SAS Insights, Data Management, ND) <https://www.sas.com/en_gb/insights/articles/data-management/gdpr-and-ai--friends--foes-or-something-in-between-.html#/> accessed 21st July 2020.

Stewart E, 'Why every website wants you to accept its cookies' (Vox Blog, 10th December 2019) <https://www.vox.com/recode/2019/12/10/18656519/what-are-cookies-website-tracking-gdpr-privacy> accessed 8th March 2021.

Stolton S, 'Artificial intelligence presents 'black swan' ethical issues, Commission report says' (Euractive Digital Blog, AI news, 8th April 2019) <https://www.euractiv.com/section/digital/news/artificial-intelligence-presents-black-swan-ethical-issues-commission-report-says/> accessed 9th April 2019.

Tanz J, 'Soon We Won't Program Computers. We'll Train Them Like Dogs' (Wired Blog, 17th May 2016) <https://www.wired.com/2016/05/the-end-of-code/> accessed 11th November 2020.

Tapson J, 'Google's Go Victory Shows AI Thinking Can Be Unpredictable, and That's a Concern' (The Conversation, 18th March 2016) <https://theconversation.com/googles-go-victory-shows-ai-thinking-can-be-unpredictable-and-thats-a-concern-56209> accessed 11th. November 2020.

Thampi A, 'Interpretable AI or How I Learned to Stop Worrying and Trust AI' (Towards Data Science, 5th March 2019) <https://towardsdatascience.com/interpretable-ai-or-how-i-learned-to-stop-worrying-and-trust-ai-e61f9e8ee2c2> accessed 15th January 2021.

Toh A, 'Rules for a New Surveillance Reality' (Human Rights Watch, 18th November 2019) <https://www.hrw.org/news/2019/11/18/rules-new-surveillance-reality> accessed 17th November 2020.

Tugrul K, 'When Even a Human is Not Good Enough as Artificial Intelligence' (Towards Data Science, 6th May 2018) <https://towardsdatascience.com/when-even-a-human-is-not-good-enough-as-artificial-intelligence-c39c9fda4644> accessed 20th May 2022.

Turner Lee N and Malamud J, 'Opportunities and blind spots in the White House's blueprint for an AI Bill of Rights' (Brookings, 19th December 2022) <https://www.brookings.edu/articles/opportunities-and-blind-spots-in-the-white-houses-blueprint-for-an-ai-bill-of-rights/> accessed 2nd October 2023.

Van Veen C, 'Landmark judgment from the Netherlands on digital welfare states and human rights' (Open Global Rights Blog, 19th March 2020) <https://www.openglobalrights.org/landmark-judgment-from-netherlands-on-digital-welfare-states/> accessed 10th December 2020.

Walker K, 'AI for Social Good in Asia Pacific' (Google post, 13th December 2018) <https://www.blog.google/around-the-globe/google-asia/ai-social-good-asia-pacific/amp/> accessed 18th November 2020.

Wallace N, 'EU's Right to Explanation: A Harmful Restriction on Artificial Intelligence' (TechZone Blog, 25th January 2017) <https://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm> accessed 5th November 2020.

Wallis P, 'Op-Ed: Counting protestors with AI changes the game forever' (Digital Journal, Technology, 8th July 2019) <https://www.digitaljournal.com/tech-science/op-ed-counting-protestors-with-a-i-changes-the-game-forever/article/553608> accessed 17th November 2020.

Wareham M, 'Stopping Killer Robots' (Human Rights Watch, 10th August 2020) <https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and> accessed 12th November 2020.

Winston E, 'GDPR – How does it impact AI?' (Information Age Blog, Data Protection and Privacy, 5th June 2023) <https://www.information-age.com/gdpr-impact-ai-123483399/> accessed 21st July 2023**.**

Wise D, 'Edward Snowden on the Dangers of Mass Surveillance and Artificial General Intelligence' (Variety Technology Blog, 26th November 2019) <https://variety.com/2019/digital/festivals/idfa-edward-snowden-1203416674/> accessed 17th November 2020.

Wycherley C, 'Uber faces landmark GDPR court challenge over alleged firing of drivers by algorithm' (GDPR Report Blog, 29th October 2020) <https://gdpr.report/news/2020/10/29/uber-faces-landmark-gdpr-court-challenge-over-alleged-firing-of-drivers-by-algorithm/> accessed 14th November 2020.

Consultation Responses:

Access Now, *Access Now's submission to the European Commission's adoption consultation on the Artificial Intelligence Act* (Feedback Reference: F2665462, August 2021).

Access Now, *Submission to the Consultation on the 'White Paper on Artificial Intelligence – a European approach to excellence and trust'* (May 2020).

Digital Europe, *Response to the European Commission's AI White Paper Consultation* (17th June 2020).

European Digital SME Alliance, *Digital SME reply to the AI Act consultation* (6th August 2021).

Future of Life Institute, *FLI Position Paper on the EU AI Act* (Feedback Reference: F2665546, August 2021).

Huawei Technologies, *Huawei response on the European Commission's Proposal for a Regulation of the European Parliament and of the Council Laying Down the Harmonised*

Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (Feedback Reference: F2665442, August 2021).

IBM, *IBM Submission to the European Commission on the Draft Artificial Intelligence Act* (Feedback Reference: F2665615, August 2021).

Insurance Europe, *Response to the consultation on draft guidelines for trustworthy artificial intelligence (AI)* (Position Paper, February 2019).

Joseph Rowntree Foundation, *Written Evidence to Parliament* (UCW0076, April 2020).

McAuley D, Joene A and Chen J, *Response to ICO consultation on the draft AI auditing framework guidance for organisations* (Horizon, 1st May 2020).

Romero Moreno F, Harbinja E, Pearce H, Mccullagh K, Sutter G, Basu S, Lysnkey O and Diker Vanberg A, *BILETA Response to UK Government Consultation: Data a new direction* (2021).

TechUK, *techUK response to the Commission's proposed Artificial Intelligence Act* (Feedback Reference: F2665579, August 2021).

University of Cambridge, Submission of Feedback to the European Commission's Proposal for a Regulation laying down harmonised rules on artificial intelligence (Feedback Reference: F2665626, August 2021).

Data Protection Authority Decisions:

Agencia Española Protección datos, 'Grindr LLC' (Spanish DPA Decision, E/03624/2021) <https://gdprhub.eu/index.php?title=AEPD_(Spain)_-_E/03624/2021> accessed 12th January 2023.

Comissão Nacional de Proteção de Dados, 'Deliberação n.º 2021/622' (Portuguese DPA Decision, 2021) <https://gdprhub.eu/index.php?title=CNPD_(Portugal)_-_Delibera%C3%A7%C3%A3o_2021/622> accessed 27th January 2023.

Datatilsynet, 'Administrative Fine – Grindr LLC' (Norwegian DPA imposes fine, 2021) <https://www.datatilsynet.no/contentassets/8ad827efefcb489ab1c7ba129609edb5/administrative-fine---grindr-llc.pdf> accessed 12th January 2023.

Datenschutz Behörde, 'Case DSB-2020.0.436.002' (Austria DPA Decision, 2020) <https://gdprhub.eu/index.php?title=DSB_(Austria)_-_2020-0.436.002> accessed 20th January 2023.

Garante per la protezione dei dati personali, 'Decision 9685994' (Italian DPA Decision, 2021) <https://gdprhub.eu/index.php?title=Garante_per_la_protezione_dei_dati_personali_(Italy)_-_9685994> accessed 28th January 2023.

Journal Articles:

Abbott K and Snidal D, ‘Hard and Soft Law in International Governance’ [2000] 54 *International Organization* 421.

Abbott R, ‘The Reasonable Computer: Disrupting the Paradigm of Tort Liability’ [2017] 86(1) *The George Washington Law Review* 101.

Ahuja A, ‘The impact of artificial intelligence in medicine on the future role of the physician’ [2019] 7 *Peer J* 1.

Aizenberg E and Van Den Hoven J, ‘Designing for Human Rights in AI’ [2020] 7(2) *Big Data & Society* 1.

Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat M and Dwivedi Y, ‘A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities’ [2023] 8(1) *Journal of Innovation & Knowledge* 100333.

Allenby B, ‘Robotics: Morals and Machines’ [2012] 481 *Nature* 26.

Aloisi A and Gramano E ‘Artificial Intelligence is watching you at work. Digital Surveillance, Employee Monitoring, and Regulatory Issues in the EU Context’ [2019] 41(1) *Comparative Labor Law and Policy Journal: Automation, Artificial Intelligence and Labour Protection* 95.

Andreotta A, Kirkham N and Rizzi M, ‘AI, Big Data, and the Future of Consent’ [2021] 37 *AI & Society* 1715.

Arrieta AB and Herrera F, ‘Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges towards Responsible AI’ [2020] 58 *Information Fusion* 82.

Asaro P, ‘On Banning Autonomous Weapons Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making’ [2012] 94 *International Review of the Red Cross* 687.

Baker A, Perov Y, Middleton K, Baxter J, Mullarkey S, Sangar D, Butt M, DoRosario A and Johri S, ‘A comparative study of Artificial Intelligence and human doctors for the purpose of triage and diagnosis’ [2020] 3 *Front. Artif. Intell.* 543405.

Baker M, ‘Promises and Platitudes: Towards a New 21st Century Paradigm for Corporate Codes of Conduct’ [2007] 23 *Connecticut Journal of International Law* 123.

Barfield W, ‘Liability for Autonomous and Artificially Intelligent Robots’ [2018] 9(1) *Journal of Behavioural Robotics* 193**.**

Bench-Capon TJM and Dunne P, ‘Argumentation in artificial intelligence’ [2007] 171(10*)* *Artificial Intelligence* 619.

Berendt B and Preibusch S, ‘Toward Accountable Discrimination-Aware Data Mining: The Important of Keeping the Human in the Loop-and Under the Looking Glass’ [2017] 5(1) *Big Data* 135.

Boddewyn J, ‘Advertising Self-Regulation: Private Government and Agent of Public Policy’ [1985] 4(1) *Journal of Public Policy & Marketing* 129.

Borgesius F, ‘Strengthening legal protection against discrimination by algorithms and artificial intelligence’ [2020] 24(10) *I.J.H.R* 1.

Brownsword R and Somsen H, ‘Law, Innovation and Technology: Before We Fast Forward- A Forum for Debate’ [2009] 1(1) *Law, Innovation and Technology* 1.

BroZek B and Jakubiec M, ‘On the legal responsibility of autonomous machines’ [2017] 25(3) *Artificial Intelligence and Law* 293.

Buccella A, ‘”AI for all” is a matter of social justice’ [2023] 3 *AI and Ethics* 11.

Buolamwini J and Gebru T ‘Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification’ [2018] 81 *Proceedings of Machine Learning Research* 1.

Casey B, Farhangi A and Vogl R, ‘Rethinking Explainable Machines: The GDPR’s ‘Right to Explanation’ Debate and the Rise of Algorithmic Audits in Enterprise’ [2019] 34 *Berkeley Technology Law Journal* 143.

Castelvecci D, ‘Beating Biometric Bias’ [2020] 587 *Nature* 347.

Cath C, ‘Governing Artificial Intelligence: Ethical, Legal, Technical Opportunities and Challenges’ [2018] 376(2133) *Philosophical Transactions of the Royal Society A.*

Cauffman C, ‘Robo-liability: The European Union in search of the best way to deal with liability for damage caused by artificial intelligence’ [2018] 25(5) *Maastricht Journal of European and Comparative Law* 527.

Čerka P, Grigienė J and Sirbikytė G, ‘Liability for damages caused by artificial intelligence’ [2015] 31(3) *Computer Law and Security Review* 376.

Chen B, Wu Z and Zhao R, ‘From fiction to fact: the growing role of generative AI in business and finance’ [2023] 21(4) *Journal of Chinese Economic and Business Studies* 471.

Chen J and Burgess P, ‘The boundaries of legal personhood: how spontaneous intelligence can problematize differences between humans, artificial intelligence, companies and animals’ [2019] 27(1) *Artificial Intelligence and Law* 73.

Chen Z, ‘Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices [2023] 10 *Humanities and Social Sciences Communications* 567.

Chesterman S, ‘Artificial Intelligence and the problem of autonomy’ [2020] 1 *Notre Dame Journal of Emerging Technologies* 210.

Chouldechova A, ‘Fair prediction with disparate impact: A study of bias in recidivism prediction instruments’ [2017] 5(2) *Big Data* 153.

Cobbe J, ‘Administrative law and the machines of government: judicial review of automated public-sector decision-making’ [2019] 39(4) *Legal Studies* 636.

Coeckelbergh M, ‘Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability’ [2020] 26 *Science and Engineering Ethics* 2051.

Coghlan S, Miller T and Paterson J, 'Good Proctor or "Big Brother"? Ethics of Online Exam Supervision Technologies [2021] 34 *Philosophy & Technology* 1581.

Courtland R, 'The bias detectives' [2018] 558 *Nature* 357.

Currie, G.M. 'Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? [2023] 53(5) *Seminars in Nuclear Medicine* 719.

Dastres R and Soori M, 'Artificial Neural Network Systems' [2021] 21(2) *I.J.I.R* 13.

Edwards L and Veale M, 'Clarity, surprises, and further questions in the Article 29 Working Part draft guidance on automated decision-making and profiling' [2018] 34(2) *Computer Law and Security Review* 398.

Edwards L and Veale M, 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?' [2018] 16(3) *IEEE Security and Privacy* 46.

Edwards L and Veale M, 'Slave to the Algorithm? Why a "Right to an Explanation" is probably not the remedy you are looking for' [2017] 16 *Duke Law and Technology Review* 18.

El-Gazzar R and Stendal K, 'Examining How GDPR Challenges Emerging Technologies' [2020] 10 *Journal of Information Policy* 237.

Etzioni A and Etzioni O, 'Should Artificial Intelligence Be Regulated?' [2017] 33(4) *Issues in Science and Technology* 32*.*

European Economic and Social Committee opinion on AI, 'The Consequence of AI on the digital single market, production, consumption, employment and society' [2017] 31(8) *Official Journal of the European Union* 1.

Felzmann H, Fosch-Villaronga E, Lutz C and Tamò-Larrieux A, 'Towards Transparency by Design for Artificial Intelligence' [2020] 26(2) *Science and Engineering Ethics* 3333*.*

Felzmann H, Fosch-Villaronga E, Lutz C and Tamo-Larrieux A, 'Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns' [2019] 6 *Big Data and Society* 1.

Gaine W, 'No-fault compensation systems' [2003] 326(7397) *BMJ* 997.

Gerstner M, 'Comment, Liability Issues with Artificial Intelligence Software' [1993] 33 *Santa Clara Law Review* 239.

Goodman B and Flaxman S, 'European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation' [2017] 38(3) *AI Magazine* 50.

Guan J, 'Artificial Intelligence in Healthcare and Medicine: Promises, Ethical Challenges and Governance' [2019] 34(2) *Chinese Medical Sciences Journal* 76.

Gugerty L, 'Newell and Simon's Logic Theorist: Historical Background and Impact on Cognitive Modelling' [2006] 50(9) *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 880.

Hagemann R, Huddleston J and Thierer A, 'Soft Law for Hard Problems: The Governance of Emerging Technologies in an Uncertain Future' [2018] 17(1) *Colorado Technology Law Journal* 37.

Hagendorff T, 'The Ethics of AI Ethics: An Evaluation of Guidelines' [2020] 30 *Minds and Machines* 99.

Haghighat M and Abdel-Mottaleb M, 'Low Resolution Face Recognition in Surveillance Systems Using Discriminant Correlation Analysis' [2017] *12th IEEE International Conference on Automatic Face and Gesture Recognition* 912.

Helberger N and Diakopoulos N, 'ChatGPT and the AI Act' [2023] 12(1) *Internet Policy Review* 1.

Hu Q, Lu Y, Pan Z, Gong Y and Yang Z, 'Can AI artifacts influence human cognition? The effects of artificial autonomy in intelligent personal assistants' [2021] 56 *International Journal of Information Management* 102250.

Hu Y, 'Robot Criminals' [2019] 52 *Michigan Journal of Law Reform* 487.

Hubbard FP, 'Sophisticated Robots' Balancing Liability, Regulation, and Innovation' [2014] 66(5) *Florida Law Review* 1803.

Hugenholtz PB and Quintais JP, 'Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?' [2021] 52 *IIC* 1190.

Jackson B, 'Artificial Intelligence and the Fog of Innovation: A Deep-Dive on Governance and the Liability of Autonomous Systems' [2019] 35(4) *Santa Clara High Technology Law Journal* 35.

Jaimes A, 'Computer vision startups tackle AI' [2016] 23(4) *IEEE* 94.

Janssen H, 'An approach for a fundamental rights impact assessment to automated decision-making' [2020] 10(1) *International Data Privacy Law* 76.

Jaume-Palasi L, 'Why Are we Failing to Understand the Societal Impact of Artificial Intelligence' [2019] 86(2) *Social Research: An International Quarterly* 477.

Jaynes T, 'Legal personhood for artificial intelligence: citizenship as the exception to the rule' [2020] 35 *AI & Society* 343.

Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong O, Shen H and Wan Y, 'Artificial intelligence in healthcare: past, present and future' [2017] 2 *Stroke and Vascular Neurology* 230.

Kalla D, Smith N, Kuraku S and Samaah F, 'Study and Analysis of Chat GPT and its Impact on Different Fields of Study' [2023] 8(3) *International Journal of Innovative Science and Research Technology* 827.

Kaptein M and Schwartz M, 'The Effectiveness of Business Codes: A Critical Examination of Existing Studies and the Development of an Integrated Research Model' [2008] 77 *Journal of Business Ethics* 111.

Karnow C, 'Liability for Distributed Artificial Intelligences' [1996] 11(1) *Berkley Technology Law Journal* 147.

Kaur J and Mann KS, 'AI based HealthCare Platform for Real Time, Predictive and Prescriptive Analytics using Reactive Programming' [2017] 933(1) *Journal of Physics: Conference Series* 1.

Kazim E, Güçlütürk O, Almeida D, Kerrigan C, Lomas E, Koshiyama A, Hilliard A and Trengove M, 'Proposed EU AI Act- Presidency compromise text: select overview and comment on the changes to the proposed regulation' [2023] 3 *AI and Ethics* 381.

Kelley R, 'Liability in Robotics: An International Perspective on Robots as Animals' [2010] 24(13) *Advanced Robotics* 1861.

Kingston J, 'Artificial Intelligence and Legal Liability' [2016] 33 *Research and Development in Intelligent Systems* 269.

Knuutila A, Herasimenko A, Au H, Bright J and Howard PN, 'A Dataset of COVID-Related Misinformation Videos and their Spread on Social Media' [2021] 7 Journal of Open Humanities Data 6.

Koulu R, 'Proceduralizing Control and Discretion: Human Oversight in Artificial Intelligence Policy [2020] 27(6) *Maastricht Journal of European and Comparative Law* 720.

Kowert W, 'The Foreseeability of Human-Artificial Intelligence Interactions' [2017] 96(1) *Texas Law Review* 181.

Kraus M, Torrez B and Hollie L, 'How Narratives of Racial Progress Create Barriers to Diversity, Equity, and Inclusion in Organizations' [2022] 43 *Current Opinion in Psychology* 108.

Kulesza T, Stumpf S, Burnett M, Yang S, Kwan I and Wong WK, 'Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' [2013] *Proceedings of IEEE Symposium on Visual Language Human-Centric Computing* 3.

Kuner C, Cate F, Lynskey O, Millard C, Loidesin N and Scantesson D, 'Expanding the Artificial Intelligence- Data Protection Debate' [2018] 8(4) *International Data Privacy Law* 289.

Land M and Aronson J, 'Human Rights and Technology: New Challenges for Justice and Accountability' [2020] 16 *Annual Review of Law and Social Science* 223.

Lanovaz M and Hranchuk K, 'Machine Learning to Analyze Single-Case Graphs: A Comparison to Visual Inspection' [2021] 54(4) *Journal of Applied Behavior Analysis* 1541.

Larsson S and Heintz F, 'Transparency in artificial intelligence' [2020] 9(2) *Internet Policy Review* 1.

Leith P, 'The rise and fall of the legal expert system' [2016] 30(3) *I.R.L.C.T* 94.

Levy D, 'Intelligent no-fault insurance for robots' [2020] 1(1) *Journal of Future Robot Life* 35

Liu H and Zawieska K, 'From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence' [2020] 22 *Ethics and Information Technology* 321.

Liu H, Lin C and Chen Y, 'Beyond State v Loomis: artificial intelligence, government algorithmization and accountability [2019] 27(2) *International Journal of Law and Information Technology* 122.

Livingstone S and Risse M, 'The Future Impact of Artificial Intelligence on Humans and Human Rights' [2019] 33(2) *Ethics and International Affairs* 141.

Loh E, 'Medicine and the rise of robotics: a qualitative review of recent advances of artificial intelligence in health [2018] 2(2) *BMJ Leader* 59.

Loideain N and Adams R, 'From Alexa to Siri and the GDPR: The gendering of Virtual Personal Assistants and the role of Data Protection Impact Assessments' [2020] 36 *Computer Law & Security Review* 105366.

Lord Griffiths, De Val P and Dormer RJ, 'Developments in English Products Liability Law: A comparison with the American system' [1988] 62 *Tulane Law Review* 354.

Lysaght T, Lim H, Xafis V and Ngiam K, 'AI-Assisted Decision-making in Healthcare' [2019] 11 *Asian Bioethics Review* 299.

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

Marciniak T, Chmielewska A, Weychan R, Parzych M and Dabrowski A, 'Influence of low resolution of images on reliability of face detection and recognition' [2015] 74 *Multimedia Tools and Applications* 4329.

Marmor A, 'Soft Law, Authoritative Advice and Non-Binding Agreements' [2019] 39(3) *Oxford Journal of Legal Studies* 507.

Massat MB, 'Artificial Intelligence in Radiology: Hype or Hope?' [2018] 47(3) Applied Radiology 22.

Matthias A, 'The responsibility gap: Ascribing responsibility for the actions of learning automa' 6(3) *Ethics and Information Technology* 175.

Mayer A, Strich F and Fiedler M, 'Unintended Consequences of Introducing AI Systems for Decision-Making' [2020] 19(4) *MIS Quarterly Executive* 239.

Mayer-Schönberger V and Range T, 'A Big Choice for Big Tech: Share Data or Suffer the Consequences' 97(5) *Foreign Affairs* 48.

Mikhaylov S, Esteve M and Campion A, 'Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration' [2018] 376(2128) *Philosophical Transaction of The Royal Society A* 1.

Mittelstadt B, 'Principles alone cannot guarantee ethical AI' [2019] 1(11) *Nature Machine Intelligence.*

Moses L, ‘How to Think about Law, Regulation and Technology: Problems with ‘Technology’ as a Regulatory Target’ [2013] 4(1) *Law, Innovation and Technology* 1.

Nemitz P, ‘Constitutional democracy and technology in the age of artificial intelligence’ [2018] 376 *Philosophical Transactions Royal Society A* 1.

Nicholson W, ‘Medical AI and Contextual Bias’ [2019] 33 *Harv. J.L. & Tech.* 66.

Obermeyer Z, Powers B, Vogeli C and Mullainathan S, ‘Dissecting racial bias in an algorithm used to manage the health of populations’ [2019] 366(6464) *Science* 447.

Oswald M, ‘Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power’ [2018] 376 *Philosophical Transactions Royal Society A* 1.

Pagallo U, ‘Apples, Orange, Robots: four misunderstandings in today’s debate on the legal status of AI systems [2018] 376(2133) *Philosophical Transactions of the Royal Society A* 1.

Pagallo U, ‘Vital, Sophia, and Co. – The Quest for the Legal Personhood of Robots’ [2018] 9(9) *Information (Switzerland)* 230.

Popescu A, ‘In Brief: Pros and Cons of Corporate Codes of Conduct [2016] 9 *Journal of Public Administration, Finance and Law* 125.

Prunkl C, ‘Human Autonomy at Risk? An Analysis of the Challenges from AI’ [2024] 34 *Minds and Machines* 26.

Reed C, ‘How Should we regulate Artificial Intelligence?’ [2018] 376(2128*) Philos Trans A Math Phys Eng Sci* 13.

Rigby M, ‘Ethical Dimensions of Using Artificial Intelligence in Health Care’ [2019] 21(2) *AMA Journal of Ethics* 121.

Risse M, ‘Human Rights and Artificial Intelligence: An Urgently Needed Agenda’ [2019] 41(1) *Human Rights Quarterly* 1.

Rodrigues R, ‘Legal and human rights issues of AI: gaps, challenges and vulnerabilities’ [2020] 4(3) *Journal of Responsible Technology* 1*.*

Romero Moreno F and Griffin JGH, ‘The UK’s criminal copyright proposals in an era of technological precision’ [2016] 7(2) *European Journal of Law and Technology* 1.

Roussi A, ‘Resisting the rise of facial recognition’ [2020] 587 *Nature* 350.

Santoni de Sio F, Almeida T and Van Den Hoven J, ‘The future of work: freedom, justice and capital in the age of artificial intelligence’ [2024] 27 *Critical Review of International Social and Political Philosophy* 659.

Schellekens M, ‘No-fault compensation schemes for self-driving vehicles’ [2018] 10(2) *Law, Innovation and Technology* 314.

Scherer M, ‘Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies’ [2016] 29(2) *Harvard Journal of Law and Technology* 354*.*

Schönberger D, 'Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications' [2019] 27(2) *International Journal of Law and Information Technology* 171.

Schuett J. 'Defining the scope of AI regulations' [2023] 15(1) *Law, Innovation and Technology* 60.

Selbst A and Powles J, 'Meaningful information and the right to explanation' [2017] 7(4) *International Data Privacy Law* 233.

Shubhendu S and Vijay J, 'Applicability of Artificial Intelligence in Different Fields of Life' [2013] 1(1) *International Journal of Scientific Engineering and Research* 2347.

Skorupinski B and Ott K, 'Technology assessment and ethics' [200] 1(2) *Poiesis & Praxis* 95.

Smuha N, 'The EU Approach to Ethics Guidelines For Trustworthy Artificial Intelligence' [2019] 20(4) *Computer Law Review International* 97.

Stănilă L, 'Artificial Intelligence and Human Rights, A Challenging Approach on the issue of equality' [2018] 2 Journal of Eastern European Criminal Law 19.

Sullivan H and Schweikart S, 'Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?' [2019] 21(2) *AMA Journal of Ethics* E160.

Tancredi L, 'Designing a no-fault alternative' [1986] 49(2) *Law and Contemporary Problems* 277.

Thomas P, Castro da Silva B, Barto A, Giguere S, Brun Y and Brunskill E, 'Preventing undesirable behaviour in intelligent machines' [2019] 366(6468) *Science* 999.

Torres P, 'The possibility and risks of artificial general intelligence' [2019] 75(3) *Bulletin of the Atomic Scientists* 105.

Truby J, 'Governing Artificial Intelligence to benefit the UN Sustainable Development Goals' [2020] 28(4) *Sustainable Development* 946.

Turing A, 'Computing Machinery and Intelligence' [1950] 59(236) *Mind* 433.

Ufert F, 'AI Regulation Through the Lens of Fundamental Rights: How Well Does the GDPR Address the Challenges Posed by AI?' [2020] 5(2) *European Papers* 1087.

Van Ooijen I and Vrabec H, 'Does the GDPR Enhance Consumers' Control over Personal Data? An Analysis from a Behavioural Perspective' [2019] 42 *Journal of Consumer Policy* 91.

Vaux J, 'From expert systems to knowledge-based companies: How the AI industry negotiated a market for knowledge' [2001] 15(3) *Social Epistemology* 231.

Veale M, 'A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence [2020] 11(1) *European Journal of Risk Regulation* 1.

Veronese A, Silveira A and Lemos A, 'Artificial Intelligence, Digital Single Market and the proposal of a right to fair and reasonable inferences: a legal issue between ethics and techniques' [2019] 5(2) *EU Law Journal* 75.

Vladeck D, ‘Machines without principals: Liability rules and artificial intelligence’ [2014] 89 *Washington Law Review* 117.

Wachter S and Mittelstadt B, ‘A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI’ [2019] 2 *Columbia Business Law Review* 1.

Wachter S, Mittelstadt B and Floridi L, ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’ [2017] 7(2) *International Data Privacy Law* 76.

Wachter S, Mittelstadt B and Russell C, ‘Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR [2018] 31(2) *Harvard Journal of Law and Technology* 841.

Weng S, Reps J, Kai J, Garibaldi J and Qureshi N, ‘Can Machine-Learning Improve Cardiovascular Risk Prediction using Routine Clinical Data? [2017] 12(4) *PLoS One* 1.

Wirtz B, Weyerer J and Geyer C, ‘Artificial Intelligence in the Public Sector- Application and Challenges’ [2018] 42(7) *International Journal of Public Administration* 596.

Witting C, ‘Duty of Care: An Analytical Approach’ [2005] 25(1) *Oxford Journal of Legal Studies* 33.

Yampolskiy R, ‘Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent’ [2020] 7(1) *Journal of Artificial Intelligence and Consciousness* 109.

Zech H, ‘Liability for AI: Public Policy Considerations’ [2021] 22 *ERA Forum* 147.

Newspaper Articles:

Conger K, Fausset R and Kovaleski S, ‘San Francisco Bans Facial Recognition Technology’ *The New York Times* (14th May 2019).

Cuthbertson A, ‘Self-Driving Cars more likely to drive into black people, study claims’ *The Independent* (6th March 2019).

Dearden L, ‘Facial recognition becoming ‘epidemic’ in British public spaces’ *The Independent* (16th August 2019).

Dearden L, ‘Information Commissioner threatens legal action against police using ‘dangerous and inaccurate’ facial recognition technology’ *The Independent* (15th May 2018).

Dodd V, ‘Cases that highlight claims of police racial profiling in England’ *The Guardian* (9th July 2020, London).

Dubal V, ‘San Francisco was right to ban facial recognition. Surveillance is a real danger’ *The Guardian* (30th May 2019).

Hern A, ‘Amazon staff listen to customers’ Alexa recordings, report says’ *The Guardian* (11th April 2019).

Hill K, ‘Wrongfully Accused by an Algorithm’ *The New York Times* (3rd August 2020).

McDonald H, ‘AI expert calls for end to UK use of ‘racially biased’ algorithms’ *The Guardian* (12th December 2019).

Mozur P, ‘One Month, 500,000 Face Scans: How China is Using AI to Profile a Minority’ *The Seattle Times* (14th April 2019).

Paris D, ‘Australia needs to face up to the dangers of facial recognition technology’ *The Guardian* (7th August 2020).

Quinn B and Perraudin F, ‘London police accused of racial profiling in lockdown searches’ *The Guardian* (16th May 2020, London).

Ram A and Neville S, ‘High-profile health app under scrutiny after doctors’ complaints’ *The Financial Times* (13th July 2018).

Ravani S, ‘Oakland committee ban on facial recognition surveillance’ *The San Francisco Chronicle* (25th June 2019).

Rawsthorn A, ‘Genius and Tragedy at Dawn of Computer Age’ *New York Times* (New York, 25 March 2012).

Roy EA, ‘New Zealand river granted same legal rights as human being’ *The Guardian* (London, 16th March 2017).

Schofield J, ‘What should I do about all the GDPR pop-ups on websites?’ *The Guardian* (5th July 2018).

Standage T, ‘Automation and Anxiety’ *The Economist* (Special Report, 25th June 2016 Edition).

Wu S, ‘Somerville City Council passes facial recognition ban’ *The Boston Globe* (27th June 2019).

Press Releases and Open Letters:

Abudu N, ‘Letter to the Orlando Police Department’ (American Civil Liberties Union, documents on use of Amazon recognition service, January 2018) <https://www.aclunc.org/docs/20180522_ARD.pdf> accessed 3rd November 2020.

Alston P, ‘Landmark ruling by Dutch court stops government attempts to spy on the poor’ (United Nations Human Rights Office of the High Commissioner, UN Special Rapporteur on extreme poverty and human rights, 5th February 2020) < https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25522> accessed 10th December 2020.

Amazon, ‘We are implementing a one-year moratorium on police use of Rekognition’ (Amazon Policy News, 10th June 2020) <https://www.aboutamazon.com/news/policy-news-

views/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition> accessed 17th November 2020.

Câmara dos Deputados, 'Chamber approves project that regulates the use of artificial intelligence' (Federal Legislative Body, News, Translated, 29th September 2021) <https://www.camara.leg.br/noticias/811702-camara-aprova-projeto-que-regulamenta-uso-da-inteligencia-artificial?utm_source=POLITICO.EU&utm_campaign=25c6120bdd-EMAIL_CAMPAIGN_2021_11_17_09_59&utm_medium=email&utm_term=0_10959edeb5-25c6120bdd-190866048> accessed 4th November 2023.

Chivot E, 'Relying on competitive advantage of AI Ethics is a losing strategy for Europe' (Center for Data Innovation Press Release, 8th April 2019) < https://datainnovation.org/2019/04/relying-on-competitive-advantage-of-ai-ethics-is-a-losing-strategy-for-europe/> accessed 18th July 2021.

Civil Liberties Union for Europe and Others, 'Open letter: The AI Act Must Protect the Rule of Law' (Open Letter, September 2023) <https://dq4n3btxmr8c9.cloudfront.net/files/iytbh9/AI_and_RoL_Open_Letter_final_27092023.pdf> accessed 1st September 2023.

Confederation of Industry of the Czech Republic, 'Open letter on the proposed regulation of artificial intelligence' (7th November 2022) <https://www.spcr.cz/images/Open_letter_on_the_proposed_regulation_of_artificial_intelligence_FIN20221107_125114.pdf> accessed 4th January 2023.

EDRi (and 44 others), *Civil Society Calls on the EU to* Prohibit Predictive and Profiling AI Systems in Law Enforcement and Criminal Justice (2022).

European Parliament, 'Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI' (European Parliament Press Release, 9th December 2023) <https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai > accessed 15th December 2023.

European Parliament, 'Artificial Intelligence in policing: safeguards needed against mass surveillance' (European Parliament Press Release, June 2021) <https://www.europarl.europa.eu/news/en/press-room/20210624IPR06917/artificial-intelligence-in-policing-safeguards-needed-against-mass-surveillance> accessed 10th January 2023.

European Parliament, 'New Product Liability Directive' (Briefing, September 2022) <https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739341/EPRS_BRI(2023)739341_EN.pdf> accessed 20th September 2023.

Good Law Project, 'Legal action over A-Level results fiasco' (News, 16th August 2020) <https://goodlawproject.org/a-level-results-fiasco/> accessed 19th September 2023.

ICO, 'ICO fines facial recognition database company Clearview AI Inc more than £7.5m and orders UK data to be deleted' (Press Release, 23rd May 2022) < https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/05/ico-fines-facial-recognition-database-company-clearview-ai-inc/> accessed 12th September 2023.

Krishna A (IBM CEO), 'Letter to Congress on Racial Justice Reform' (IBM, 11th November 2020) <https://www.ibm.com/policy/facial-recognition-sunset-racial-justice-reforms/> accessed 18th November 2020.

Liberty, 'Liberty wins ground-breaking victory against facial recognition tech' (Press Release, 11th August 2020) <https://www.libertyhumanrights.org.uk/issue/liberty-wins-ground-breaking-victory-against-facial-recognition-tech/> accessed 5th November 2020.

Mayor P, 'Letter to Chief Investigator' (American Civil Liberties Union, 24th June 2020) <https://www.aclu.org/letter/aclu-michigan-complaint-re-use-facial-recognition> accessed 18th November 2020.

Office of Qualifications and Examinations Regulation (Ofqual), 'Statement from Roger Taylor, Chair, Ofqual' (Press Release, 17th August 2020) <https://www.gov.uk/government/news/statement-from-roger-taylor-chair-ofqual> accessed 19th September 2023.

Robonarratives, 'My response to the EU's intentions to grant robots 'Electronic Personhood' and the Open Letter to the European Commission' (By a scientist that signed the Open Letter, 19th April 2018) <https://robonarratives.wordpress.com/2018/04/19/my-response-to-eus-intentions-to-grant-robots-electronic-personhood-and-the-open-letter-to-the-european-commission/> accessed 29th March 2019.

Robotics Openletter, 'Open Letter to the European Commission, Artificial Intelligence and Robotics' (2018) <https://robotics-openletter.eu/#:~:text=We%2C%20Artificial%20Intelligence%20and%20Robotics,Union%20citizens%20while%20fostering%20innovation.> accessed 10th June 2021.

Sterken F, 'AI isn't 100%' (Indica, 28th June 2022) < https://indica.nl/blog/2022/28/6/ai-artificial-intelligence> accessed 29th June 2022.

The Surveillance Camera Commissioner, 'statement on Court of Appeal judgment (R) Bridges v South Wales Police – Automated Facial Recognition' (Gov.uk, press release, 11th August 2020) <https://www.gov.uk/government/speeches/surveillance-camera-commissioners-statement-court-of-appeal-judgment-r-bridges-v-south-wales-police-automated-facial-recognition> accessed 5th November 2020.

Volvo, 'US urged to establish nationwide Federal guidelines for autonomous driving' (Press Release, 7th October 2015) <https://www.media.volvocars.com/global/en-gb/media/pressreleases/167975/us-urged-to-establish-nationwide-federal-guidelines-for-autonomous-driving> 11th November 2020.

Wayne County Prosecuting Office, 'Statement in response to New York Times Article Wrongfully Accused by an Algorithm' (Press Release, 24th June 2020) < https://int.nyt.com/data/documenthelper/7046-facial-recognition-arrest/5a6d6d0047295fad363b/optimized/full.pdf#page=1> accessed 19th November 2020.

White House, 'ICYMI:Wired (Opinion): Americans Need a Bill of Rights for an AI-Powered World' (White House News and Updates, 22nd October 2021) <https://www.whitehouse.gov/ostp/news-updates/2021/10/22/icymi-wired-opinion-americans-need-a-bill-of-rights-for-an-ai-powered-world/> accessed 1st October 2023.

Reports and Publications:

Access Now, *Human Rights in the Age of Artificial Intelligence* (2018).

Access Now, *The European Human Rights Agenda in the Digital Age* (February 2020).

AI Now, *AI Now Report* (2019).

AI Now, *Disability, Bias and AI* (2019).

Amnesty International and Access Now, *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems* (2018).

Article 29 Working Party, *WP251 Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679* (February 2018).

Axon AI and Policing Technology Ethics Board, *First Report of the Axon AI & Policing Technology Ethics Board* (2019).

Bogucki A, Engler A, Perarnaud C, and Renda A, *The AI Act and Emerging EU Digital Acquis: Overlaps, gaps and inconsistencies* (CEPS, September 2022).

Brave, *DPA Report* (2020).

Cambridge Consultants, *Use of AI in Online Content* Moderation (on behalf of Ofcom, 2019).

Campolo A, Sanfilippo M, Whittaker M and Crawford K, *AI Now 2017 Report* (2017).

Castro D, *Benefits and Limitations of Industry Self-Regulation for Online Behavioral Advertising* (The Information Technology and Innovation Foundation, 2011).

Center for Data Innovation, *How Much Will the Artificial Intelligence Act Cost Europe?* (July 2021)

Center for Data Innovation, *Recommendations to the EU High Level Expert Group on Artificial Guidelines for Trustworthy AI* (2019).

Centre for Data Ethics and Innovation, *Interim Report: Review into bias in algorithmic decision-making* (July 2019).

Centre for Data Ethics and Innovation, *Review into bias in algorithmic decision-making* (2020).

Centre for Information Policy Leadership, *Artificial Intelligence and Data Protection – How the GDPR Regulates AI* (2020).

Centre for Information Policy Leadership, *Artificial intelligence and Data Protection: Delivering Sustainable AI Accountability in Practice, First Report: Artificial Intelligence and Data Protection in Tension* (10th October 2018).

Commissioner for Human Rights, *Unboxing Artificial Intelligence: 10 steps to protect Human Rights – Recommendation* (Council of Europe, 2019).

Congressional Research Service, *Highlights of the 2023 Executive Order on Artificial Intelligence for Congress* (November 2023).

Council of Europe, *Report on the Rule of Law* (Venice Commission, 2011).

Datenethikcommission, *Opinion of the Data Ethics Commission*. (Data Ethics Commission, German Federal Ministry of Justice and Consumer Protection, 2019).

Davies B, Innes M and Dawson A, *An Evaluation of South Wales Police's Use of Automated Facial Recognition* (Cardiff University, September 2018).

Deloitte, *Transparency and Responsibility: A call for explainable AI* (2019).

Department for Digital, Culture, Media & Sport, and the Home Office, *Consultation outcome, Online Harms White Paper* (updated December 2020).

Department for Digital, Culture, Media & Sport, *Data: A New Direction- Government Response to Consultation* (June 2022).

Department for Science, Innovation and Technology, *White Paper: A pro-innovation approach to AI regulation* (March 2023, updated February 2024).

Department of Health and Social Care, *A guide to good practice for digital and data-driven health technologies* (Government Department, 2021).

Dignum V, Muller C and Theodorou A, *Final Analysis of the EU Whitepaper on AI* (ALLAI, June 2020).

EDRi, *Ban Biometric Mass Surveillance* (2020).

EDRi, *The Rise and Rise of Biometric Mass Surveillance in the EU* (2021).

Edwards L, *The EU AI Act: a summary of its significance and scope* (Ada Lovelace Institute, April 2022).

Engler A, *The Case for AI Transparency Requirements* (The Brookings Institution's Artificial Intelligence and Emerging Technology Initiative, January 2020).

Equality and Human Rights Commission, *Data Protection and Digital Information Bill* (House of Lords, 2nd Reading, 15/12/2023).

Equality and Human Rights Commission, *Stop and think – A critical review of the use of stop and search powers in England and Wales* (2010).

European Commission, *Annexes to the 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- Fostering a European approach to Artificial Intelligence, COM (2021)

European Commission, *Civil Liability – adapting liability rules to the digital age and artificial intelligence* (Inception Impact Assessment, 2021-22).

European Commission, *Commission Decision of 24.1.2024 establishing the European Artificial Intelligence Office* (390 final, 2024).

European Commission, *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- Artificial Intelligence for Europe,* COM (2018).

European Commission, *Draft Ethics Guidelines for Trustworthy AI* (High-Level Expert Group on Artificial Intelligence, 18th December 2018).

European Commission, *Liability for Artificial Intelligence and other emerging digital technologies* (Expert Group on Liability and New Technologies- New Technologies Formation, 2019).

European Commission, *Proposal for a Directive of the European Parliament and of the Council on liability for defective products COM 2022/0302(COD)* (495 final, 2022).

European Commission, *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to Artificial Intelligence (AI Liability Directive)*, COM (2022).

European Commission, *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts*, COM (2021)

European Commission, *Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC* (COM 2020 825).

European Commission, *Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee- Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics,* COM (2020).

European Commission, *Science for Environment Policy, Future Brief: The Precautionary Principle: Decision-Making under Uncertainty* (Issue 18, 2017).

European Commission, *Structure for the White Paper on Artificial Intelligence – a European Approach* (2020, Draft 12/12).

European Commission, *White Paper on Artificial Intelligence – A European approach to excellence and trust,* COM (2020).

European Court of Human Rights, *Article 7: The quality of law requirements and principle of non-retrospectiveness of the criminal law under Article 7 of the Convention* (2019, Council of Europe Research Division).

European Court of Human Rights, *Guide on Article 8 of the European Convention on Human Rights* (updated 31st August 2020).

European Data Protection Board, *EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act)* (2021).

European Data Protection Board, *Guidelines 05/2020 on consent under Regulation 2016/679* (2020).

European Data Protection Supervisor, *EDPS Opinion on the European Commission's White Paper on Artificial Intelligence – A European approach to excellence and trust* (2020).

European Data Protection Supervisor, *Orientations from the EDPS. Reactions of EU Institutions as Employers to the COVID-19 Crisis* (2020).

European Data Protection Supervisor, *TechDispatch: Explainable Artificial Intelligence* (2023).

European Parliament, *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))* (June 2023).

European Parliament, *Artificial Intelligence and Civil Liability* (Policy Department for Citizens' Rights and Constitutional Affairs, 2020).

European Parliament, *Artificial Intelligence in transport: Current and future developments, opportunities and challenges* (European Parliamentary Research Service, 2019).

European Parliament, *Artificial Intelligence Liability Directive* (Briefing, EU Regulation in Progress, February 2023).

European Parliament, *Draft Opinion of the Committee on Industry, Research and Energy* (Rapporteur for opinion: Eva Maydell, March 2022).

European Parliament, *Draft Report on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM2021/0206 – C9-0146/2021 – 2021/0106(COD))* (2022).

European Parliament, *Provisional Agreement Resulting from Interinstitutional Negotiations – Proposal for a regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts 2021/0106 (COD)* (2019-2024).

European Parliament, *Report with recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103(INL)).

European Parliament, *Resolution of 20th October 2020 with recommendations to the Commission on a civil liability regime for Artificial Intelligence (2020/2014(INL))* (2021).

European Parliament, *Resolution of 3 May 2022 on artificial intelligence in a digital age (2020/2266(INI))* (2022).

European Parliament, *Resolution on Fundamental Rights Implications of Big Data: Privacy, Data Protection, Non-Discrimination, Security and Law-Enforcement* (2017, 2016/2225(INI)).

European Parliament, *The ethics of artificial intelligence: Issues and initiatives* (European Parliamentary Research Service, March 2020).

European Parliament, *The Impact of the General Data Protection Regulation (GDPR) on artificial intelligence* (European Parliamentary Research Service, June 2020).

European Parliament, *Understanding algorithmic decision-making: Opportunities and challenges* (2019).

European Union Agency for Fundamental Rights, *#BigData: Discrimination in data-supported decision making* (2018, FRA Focus).

European Union Agency for Fundamental Rights, *Facial Recognition Technology: Fundamental rights considerations in the context of law enforcement* (2019).

European Union Agency for Fundamental Rights, *Getting the Future Right, Artificial Intelligence and Fundamental Rights* (2020).

Fair Trials, *Automating Injustice* (2021).

Floridi L, Holweg M, Taddeo M, Silva JA, Mökander J and Wen Y, *capAI- A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act* (Version 1.0, March 2022).

Fussey P and Murray D, *Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology* (The Human Rights, Big Data and Technology Project, 2019).

Future of Humanity Institute, University of Oxford, Centre for the Study for Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation and Open AI, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation* (2018).

Future of Privacy Forum, *Automated Decision-Making Under the GDPR* (May 2022).

Future of Privacy Forum, *Unfairness by Algorithm: Distilling the harms of automated decision-making* (2017).

Government of Canada, *The Artificial Intelligence and Data Act – Companion Document* (2022).

High-Level Expert Group on Artificial Intelligence, *A Definition of AI: Main Capabilities and Disciplines* (Independent group set up by the European Commission, 2019).

High-Level Expert Group on Artificial Intelligence, *Guidelines for Trustworthy AI* (Independent group set up by the European Commission, 8th April 2019).

House of Lords, *Order of Business* (Volume 794, No. 208, Select Committee Report on Artificial Intelligence, 19th November 2018).

House of Lords, *Universal Credit isn't working: proposals for reform'* (2nd Report of Session 2019-21, Economic Affairs Committee, published July 2020).

Human Rights Watch, *Automated Hardship* (September 2020).

Information Commissioner's Office, *Big Data, Artificial Intelligence, Machine Learning and Data Protection* (2017, Version 2.2).

Information Commissioner's Office, *Data Protection at the end of the transition period* (2019)

Information Commissioner's Office, *Guidance on AI and Data Protection* (2020, updated 2023).

Information Commissioner's Office, *Guidance on automated decision-making and profiling* (June 2018).

Information Commissioner's Office, *Guidance on Lawful basis for processing: Consent'* (March 2018).

Information Commissioner's Office, *Guidance on the AI auditing framework* (2020).

Information Commissioner's Office, *Guide to the General Data Protection Regulation (GDPR)* (2018, last updated: March 2022).

Information Commissioner's Office, *How do we ensure individual rights in our AI systems?* (July 2020).

Information Commissioner's Office, *Project ExplAIn Interim Report* (2019).

Information Commissioner's Office, *Rights Related to automated decision making including profiling* (July 2020).

Information Commissioner's Office, *Summary of response to the consultation on ICO guidance on the AI auditing framework, with comments* (2020).

Information Commissioner's Office, *The use of live facial recognition technology by law enforcement in public places* (2019).

Information Commissioner's Office*, The use of live facial recognition technology in public places* (2021).

Innovation, Science and Economic Development Canada, *Canada's Digital Charther in Action: A Plan by Canadians, for Canadians* (2019).

Latonero M, *Governing Artificial Intelligence: Upholding Human Rights & Dignity* (2018, Data & Society).

MacCarthy M, *AI needs more regulation, not less* (on behalf of The Brookings Institution's Artificial Intelligence and Emerging Technology initiative, 2020).

Marchal N, Kollanyi B, Neudert L, Au H and Howard PN, *Junk News & Information Sharing During the 2019 UK General Election* (Data Memo 2019.4, University of Oxford, 2019).

National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (US Department of Commerce, January 2023).

National Institute of Standards and Technology, *Facial Recognition Vendor Test (FRVT) Part 3: Demographic Effects* (2019).

OECD, *Recommendation of the Council on Artificial Intelligence* (Legal Instrument 0049, adopted May 2019, Amended November 2023).

Office for Product Safety & Standards, *Code of Practice on consumer product safety related recalls and other corrective actions* (PAS 7100:2018).

Privacy International and Article 19, *Privacy and Freedom of Expression in the Age of Artificial* Intelligence (April 2018).

Privacy International, *Data is Power: Profiling and Automated Decision-Making in* GDPR (April 2018).

Raso F, Hilligoss H, Krishnamurthy V, Bavitz C and Kim L, *Artificial Intelligence and Human Rights: Opportunities and Risks* (Berkman Klein Center Research Publication, 2018).

RECIPES Project, *Intra Case Study Analysis* (Funded by EU's Horizon Research, 2020).

Renda A, *Beyond the Brussels Effect* (Policy Brief, Foundation for European Progressive Studies, March 2022).

Royal Academy of Engineering, *Autonomous Systems: Social, Legal and Ethical Issues* (2009).

Science and Technology Committee, *Robotics and Artificial Intelligence* (Fifth Report of Session 2016-2017, House of Commons, 2017).

Silberg J and Manyika J, *Notes from the AI frontier: Tackling bias in artificial intelligence (and in humans)* (McKinsey Global Institute, 2019).

The Committee on Standards in Public Life, *Artificial Intelligence and Public Standards* (February 2020, Independent Report).

The Norwegian Data Protection Authority, *Artificial Intelligence and Privacy* (January 2018).

The Royal Society, *Explainable AI: The Basics* (Policy Briefing, November 2019).

Tucker J and Norris D, *Rough Justice* (2018, Child Poverty Action Group).

UK Government, *National AI Strategy* (September 2021).

UK Government, *Policy Paper- Establishing a pro-innovation approach to regulating AI* (July 2022).

UNESCO's COMEST, *Report of COMEST on Robotic Ethics* (Paris, September 2017).

United Nations Human Rights Committee, *General Comment No.16: Article 17 (The right to respect of privacy, family, home and correspondence, and protection of honour and reputation)* (Thirty-second Session, 8th April 1988).

United Nations, *Extreme Poverty and Human Rights* (Seventy-fourth session, 11th October 2019).

United Nations, *Promotion and protection of the right to freedom of opinion and expression* (August 2018, General Assembly, Seventy-Third Session).

Wendehorst C, *AI Liability in Europe: anticipating the EU AI Liability Directive* (Ada Lovelace Institute, September 2022).

West D and Allen J, *How Artificial Intelligence is Transforming the World* (Brookings EDU Report, 2018).

White House, *Blueprint for an AI Bill of Rights* (October 2022).

White House, *Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence* (October 2023).

Yaros O, Hajda O, Prinsley MA, Randall R and Hepworth E, *UK Government Proposes a New Approach to Regulating Artificial Intelligence (AI)* (Mayer Brown, August 2022).

Textbooks:

Bygrave LA, *The EU General Data Protection Regulation (GDPR) – A Commentary*, (2020, 1st edition, Oxford University Press).

Castrounis A, *AI for People and Business: A Framework for Better Human Experiences and Business Success* (O'Reilly Media, 2019).

Chopra S and White L, *A Legal Theory for Autonomous Artificial Agents* (University of Michigan Press, 2011).

Chrisley R, *Artificial Intelligence: Critical Concepts, Volume 1* (Taylor & Francis, 2000).

Coeckelbergh M, *AI Ethics* (The MIT Press Essential Knowledge Series, 2020).

Donnelly, J. *Universal Human Rights in Theory and Practice* (Cornell University Press, 2013).

Downes L, *The Laws of Disruption: Harnessing the New Forces that Govern Life and Business in the Digital Age* (Business & Economics, 2009).

Goodhart C, *Financial Regulation: Why, how and where now?* (Routledge, 1998).

Haufler V, *A Public Role for the Private Sector: Industry Self-Regulation in a Global Economy* (Carnegie Endowment for International Peace, 2001).

Howard M, *Artificial Intelligence, Machine Learning and Deep Learning,* (CreateSpace Publishing, 2018).

Kaplan J, *Artificial Intelligence: What Everyone Needs to Know* (Oxford University Press, 2016).

Loukides M and Lorica B, *What is Artificial Intelligence?* (O'Reilly Publishing, 2016).

Osaba O and Welser W, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence* (Rand Corporation, 2017).

Senden L, *Soft law in European Community Law* (2004, Hart Publishing).

Shelton D, *Commitment and Compliance: The Role of Non-Binding Norms in the International Legal System* (Oxford University Press, 2000).

Taulii T, *Artificial Intelligence Basics: A Non-Technical Introduction* (Apress Publishing, 2019).

Turner J, *Robot Rules – Regulating Artificial Intelligence* (Palgrave Macmillan, 2019).

Textbook Chapters:

Asaro P, 'A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics' in Lin P, Abney K and Bekey G (eds), *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press, 2012).

Chrysafiadi K, 'The Role of Fuzzy Logic in Artificial Intelligence and Smart Applications' in Tsihrintzis GA, Virvou M and Jain LC (eds), *Learning and Analytics in Intelligent Systems* (Springer, Cham, Volume 34, 2023).

Clifford D, Richardson M and Witzleb N, 'Artificial Intelligence and Sensitive Inferences: New Challenges for Data Protection Laws' in Findlay M, Ford J, Seoh J and Thampapillai D (eds), *Regulatory Insights on Artificial Intelligence: Research for Policy* (2021, Edward Elgard).

Haenold S, 'Profiling and Automated Decision-Making: Legal Implications and Shortcomings' in Corrales M, Fenwick M and Forgó N (eds), *Robotics, AI and the Future of Law* (2018).

Mendoza I and Bygrave LA, 'The Right Not to Be Subject to Automated Decisions Based on Profiling' in Synodinou T, Jougleux P, Markou C and Prastitou T (eds), *EU Internet Law: Regulation and Enforcement* (Springer, 2017).

Mohallick I, De Moor K, Özgöbek Ö and Gulla J, 'Towards New Privacy Regulations in Europe: Users' Privacy Perception in Recommender Systems' in Wang G, Chen J and Yang L (eds), *Security, Privacy and Anonymity in Computation, Communication, and Storage* (2018, Springer).

Salem M, Lakatos G, Amirabdollahian F and Dautenhahn K, 'Towards safe and trustworthy social robots: ethical challenges and practical issues' in Tapus A, André E, Martin JC, Ferland F, Ammi M (eds), *Social Robotics*. (ICSR, Lecture Notes in Computer Science, Springer, Cham, Vol. 9388, 2015).

Wagner B, '*Ethics as an escape from regulation'* in Bayamlioğlu E, Baraliuc I, Janssens L (eds), *Being Profiled* (Amsterdam University Press, 2018).

Wrigley S, 'Taming Artificial Intelligence: "Bots," the GDPR and Regulatory Approaches' in Corrales M, Fenwick M, and Forgó N (eds), *Robotics, AI and the Future of Law. Perspectives in Law, Business and Innovation* (Springer Singapore, 2018).

Websites:

Amazon, 'What is Artificial Intelligence? Machine Learning and Deep Learning' (Amazon website) <www.aws.amazon.com/machine-learning/what-is-ai/> accessed 12th July 2019.

Autonomous Weapons, 'Ban Lethal Autonomous Weapons' (access to video and pledge) <https://autonomousweapons.org/> accessed 12th November 2020.

Ban Facial Recognition, 'Ban Facial Recognition' (Website, Interactive Map) <https://www.banfacialrecognition.com/map/> accessed 2nd October 2023.

Big Brother Watch, 'Active Campaigns' (Pressure Group Website) <https://bigbrotherwatch.org.uk/campaigns/ accessed 3rd January 2022.

British Council, 'Should robots be citizens?' (British Council website) <https://www.britishcouncil.org/anyone-anywhere/explore/digital-identities/robots-citizens> accessed 14th November 2020.

Department of Health and Social Care, 'New Code of Conduct for AI systems used by the NHS (Research and Innovation, Government website, 19th February 2019) <https://www.gov.uk/government/news/new-code-of-conduct-for-artificial-intelligence-ai-systems-used-by-the-nhs> accessed 5th March 2021.

Digibyte, 'European Commission publishes ranking guidelines under the P2B Regulation to increase transparency of online search results' (December 2020, EU Website) <https://digital-strategy.ec.europa.eu/en/news/european-commission-publishes-ranking-guidelines-under-p2b-regulation-increase-transparency-online> accessed 6th January 2022.

EDRi, 'Remote Biometric Identification: a technical and legal guide' (EDRi website, 23rd January 2023) <https://edri.org/our-work/remote-biometric-identification-a-technical-legal-guide/> accessed 11th September 2023.

Entilic, 'Intelligence that cares' (Corporation Website) <https://www.enlitic.com/> accessed 2nd January 2022.

Ethics for Artificial Intelligence, 'Problems with Codes of Ethics' (website) <https://www.cs.ox.ac.uk/efai/developing-codes-of-ethics-for-ai/downsides-of-codes-of-ethics/> accessed 5th February 2020.

European Data Protection Supervisor, 'Artificial Intelligence' (EDPS website) <https://edps.europa.eu/data-protection/our-work/subjects/artificial-intelligence_en> accessed 15th November 2020.

European Parliament Think Tank, 'The Precautionary Principle: Definitions, Applications and Governance' (2015, European Parliament) <https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_IDA(2015)573876> accessed 4th December 2019.

Google Cloud, 'Explainable AI' (Corporation Website) <https://cloud.google.com/explainable-ai > accessed 12th January 2022.

Google, ‘Artificial Intelligence at Google: Our Principles’ (official website) <https://ai.google/principles/> accessed 18th November 2020.

Government of Canada, ‘Algorithmic Impact Assessment’ (Canadian Government Website, Algorithm Impact Tool, 2019) <https://canada-ca.github.io/aia-eia-js/> accessed 22nd January 2021.

Government of Canada, ‘Guide on the use of generative artificial intelligence’ (Canadian Gov Website) <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html> accessed 4th September 2024.

Government of Canada, ‘Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems’ (2023, Canadian Gov Website) <https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems> accessed 4th September 2024.

IBM, ‘AI Fairness 360’ (2018, Corporation Website) <https://aif360.res.ibm.com/> accessed 5th January 2022.

IBM, ‘IBM’s Principles for Trust and Transparency’ (IBM website) <https://www.ibm.com/policy/wp-content/uploads/2018/06/IBM_Principles_SHORT.V4.3.pdf> accessed 18th November 2020.

IBM, ‘Shedding light on AI bias with real world examples’ (IBM website) <https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples> accessed 4th September 2024.

IBM, ‘Trustworthy AI; (IBM website) <www.research.ibm.com/5-in-5/ai-and-bias/> accessed 11th June 2019.

IEEE, ‘Algorithmic Bias Considerations’ (2017, P7003 Project) < https://standards.ieee.org/ieee/7003/6980/> accessed 10th January 2022.

Information Commissioner’s Office, ‘Special Category Data’ (ICO Website and Guidance) <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-is-special-category-data/#scd7> accessed 2nd March 2021.

Information Commissioner’s Office, ‘Taking your case to court and claiming compensation’ (ICO Website) <https://ico.org.uk/for-the-public/data-protection-and-journalism/taking-your-case-to-court-and-claiming-compensation/> accessed 20th January 2023.

Information Commissioners Office, ‘Who we are’ (ICO Website) <https://ico.org.uk/about-the-ico/who-we-are/> accessed 15th November 2020.

Liberty, ‘Legal Action’ (Pressure Group Website) <https://www.libertyhumanrights.org.uk/?s=legal+action> accessed 3rd January 2022.

Metropolitan Police, 'Gangs Violence Matrix' (Met. Police website) <https://www.met.police.uk/police-forces/metropolitan-police/areas/about-us/about-the-met/gangs-violence-matrix/> accessed 20th August 2023.

Microsoft, 'Principles and Approach (Official website) <https://www.microsoft.com/en-us/ai/principles-and-approach/> accessed 18th November 2020.

New York Government Website, 'New York City Automated Decisions Systems Task Force' (New York, USA) <https://www1.nyc.gov/site/adstaskforce/index.page> accessed 17th November 2020.

Office of Qualifications and Examinations Regulation (Ofqual), 'About us' (Government website) <https://www.gov.uk/government/organisations/ofqual/about> accessed 19th September 2023.

Ofsted, 'About Us' (Government Website) <https://www.gov.uk/government/organisations/ofsted/about> accessed 15th June 2023.

Partnership on AI, 'About Us' (official website) < https://partnershiponai.org/about/> accessed 4th February 2020.

Partnership on AI, 'Meet the Partners' (official website) <https://www.partnershiponai.org/partners/> accessed 4th February 2020.

Reclaim Your Face, 'Reclaim Your Face' (Website) <https://reclaimyourface.eu/> accessed 13th September 2023.

SAS, 'Machine Learning, what it is and why it matters' (SAS Analytics and Data Science Insights) <www.sas.com/en_gb/insights/analytics/machine-learning.html> accessed 15th July 2019.

TechUK, 'Government proposals for UK AI regulation' (TechUK Website, 18th July 2022) <https://www.techuk.org/resource/government-proposals-for-uk-ai-regulation.html> accessed 26th September 2023.

The Alan Turing Institute, 'A right to explanation' (Advice from Turing researchers) <https://www.turing.ac.uk/research/impact-stories/a-right-to-explanation> accessed 4th March 2020.

The Human Rights, Big Data, and Technology Project, 'About Us' <https://www.hrbdt.ac.uk/about-us/> accessed 17th November 2020.

United Nations, 'Human Rights' (UN Website) <https://www.un.org/en/global-issues/human-rights> accessed 4th September 2024.

Working Papers and Conference Papers:

Asaro P, 'Robots and Responsibility from a Legal Perspective' (IEEE Conference on robotics and automation, Robo-ethics, Rome, 2007)

<https://www.peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf> accessed 20th January 2021.

Benhamou Y and Ferland J, 'Artificial Intelligence and Damages: Assessing Liability and Calculating the Damages' (Leading Legal Disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law, 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3535387> accessed 20th June 2021.

Chu C, Zhmoginov A and Sandler M, 'CycleGAN, a Master of Steganography' (NIPS Workshop 'Machine Deception', 2017) <https://storage.googleapis.com/gweb-research2023-media/pubtools/pdf/d8511d26602659849b93d28875f25780e37a973d.pdf> accessed 12th June 2020.

Cummings M, 'Artificial Intelligence and the Future of Warfare' (International Security Department and US and the Americas Programme, 2017) <https://www.chathamhouse.org/sites/default/files/publications/research/2017-01-26-artificial-intelligence-future-warfare-cummings-final.pdf> accessed 20th March 2020**.**

Danks D and London A, 'Algorithmic Bias in Autonomous Systems' (Proceedings of the 26th International Joint Conference on Artificial Intelligence 2017) <https://www.researchgate.net/publication/318830422_Algorithmic_Bias_in_Autonomous_Systems> accessed 21st March 2020).

Dhurandhar A, Iyengar V, Luss R and Shanmugam K, 'A Formal Framework to Characterize Interpretability of Procedures' (ICML Workshop on Human Interpretability, Sydney, 2017) <https://arxiv.org/abs/1707.03886> accessed 19th March 2022.

Dignum V, 'Responsible Autonomy' (Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017) <https://arxiv.org/pdf/1706.02513.pdf> accessed 19th June 2023).

Doshi-Velez F and Kim B, 'Towards a Rigorous Science of Interpretable Machine Learning' (2017) <https://arxiv.org/abs/1702.08608> accessed 15th January 2023.

Doshi-Velez F, Kortz M, Budish R, Bavitz C, Gershman S, O'Brien D, Scott K, Schieber S, Waldo J, Weinberger D and Wood A 'Accountability of AI Under the Law: The Role of Explanation' (Berkman Center Research Publication, 2017, revised 2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3064761> accessed 12th January 2023.

Feldstein S, 'The Global Expansion of AI Surveillance' (Carnegie Endowment for International Peace, 2019) <https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.pdf> accessed 27th July 2021.

Fjeld J, Achten N, Hilligoss H, Nagy A and Srikumar M, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' (Berkman Klein Center for Internet and Society at Harvard University, 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482> accessed 22nd July 2021**.**

Gaviria C, 'The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps Generated by AI in the US' (Pardee Rand, 2020) <https://www.rand.org/pubs/rgs_dissertations/RGSDA319-1.html> accessed 10th August 2021.

Gilpin L, Yuan B, Bajwa A, Specter M and Kagal L, 'Explaining Explanations: An Overview of Interpretability of Machine Learning' (IEEE 5th International Conference on Data Science and Advanced Analytics, 2018) <https://arxiv.org/abs/1806.00069> accessed 20th July 2021.

Hamid S, 'The Opportunities and Risks of Artificial Intelligence in Medicine and Healthcare' (University of Cambridge, 2016) <https://api.repository.cam.ac.uk/server/api/core/bitstreams/d4b6cb45-f7fc-45bd-bcd2-679801cefbe0/content> accessed 15th August 2022.

Hind M, Wei D, Campbell M, Codella NCF, Dhurandhar A, Mojsilovic A, Ramamurthy KN and Varshney KR, 'TED: Teaching AI to Explain its Decisions' (IBM Research, 2019) <https://arxiv.org/pdf/1811.04896.pdf> accessed 12th January 2023.

Hussain F, Hussain R and Hossain E, 'Explainable Artificial Intelligence (XAI): An Engineering Perspective' (January 2021) <https://www.semanticscholar.org/reader/1f0d09386ee7685c4a8953aed81adbd4055763c1> accessed 4th September 2024.

Jannai D, Meron A, Lenz B, Levine Y and Shoham Y 'Human or Not? A Gamified Approach to the Turing Test' (2023) <https://arxiv.org/pdf/2305.20010.pdf> accessed 30th July 2023.

Janßen R, Kesler R, Kummer ME and Waldfogel J, 'GDPR and the Lost Generation of Innovative Apps' (NBER Working Paper No. w30028, July 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4104014> accessed 1st August 2023.

Kazim E and Koshiyama A, 'A review of the ICO's Draft Guidance on the AI Auditing Framework' (SSRN, 2020) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3599226> accessed 23rd July 2022.

Kleinberg J, Mullainathan S and Raghavan M, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' (Cornell University, 17th November 2016) < https://arxiv.org/abs/1609.05807v2> accessed 5th November 2020.

Langer PF, 'Lessons from China – The Formation of a Social Credit System: Profiling, Reputation Scoring, Social Engineering' (The 21st Annual International Conference on Digital Government Research, pages 164-174, June 2020) <https://dl.acm.org/doi/10.1145/3396956.3396962> accessed 5th September 2023.

McCarthy J, Minsky ML, Rochester N and Shannon CE, 'A Proposal For the Dartmouth Summer Research Project on Artificial Intelligence' (1955) <https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html> accessed 12th February 2020.

Olson Jr CE, 'Is 80% accuracy good enough?' (Senior Image Analyst and Michigan Tech Research Institute, 2008) <https://www.asprs.org/a/publications/proceedings/pecora17/0026.pdf> accessed 13th February 2019.

Poursabzi-Sangdeh F, Goldstein D, Hofman J, Vaughan J and Wallach H, 'Manipulating and measuring model interpretability' (Cornell University, 8th November 2020) <https://arxiv.org/abs/1802.07810> accessed 6th November 2020.

Pownall C, 'Understanding the reputational risks of AI' (CPC & Associates, AI Trust & Transparency Project, 2019) <https://www.researchgate.net/publication/340088726_Understanding_the_Reputational_Risks_of_AI> accessed 20th March 2021.

Price II WN, 'Artificial Intelligence in Health Care: Applications and Legal Implications' (University of Michigan, 2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3078704> accessed 16th March 2020.

Sunstein C, 'Beyond the Precautionary Principle' (University of Chicago, 2002) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=307098> accessed 1st July 2022.

Tarelli E, 'The Strengths and Weaknesses of Soft Law as a Source of International Financial Regulation' (SSRN, 2009) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1467842> accessed 19th February 2020.

Verma S and Rubin J, 'Fairness Definitions Explained' (International Workshop on Software Fairness, Sweden, Fairware, 2018) <https://fairware.cs.umass.edu/papers/Verma.pdf> accessed 12th August 2022.

Whittlestone J, Nyrup R, Alexandrova A and Cave S, 'The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions' (University of Cambridge, 2019) <http://lcfi.ac.uk/media/uploads/files/AIES-19_paper_188_Whittlestone_Nyrup_Alexandrova_Cave.pdf> accessed 13th March 2021.

Yampolskiy R, 'Unpredictability of AI' (2019) <https://arxiv.org/ftp/arxiv/papers/1905/1905.13053.pdf> accessed 15th April 2021.