algo:aware

Raising awareness on algorithms

Procured by the European Commission's Directorate-General for Communications Networks, Content and Technology

State-of-the-Art Report | Algorithmic decision-making

Version 1.0

December 2018

Table of Contents

Executive Summary

Context

Algorithmic systems are present in all aspects of modern lives. They are sometimes involved in mundane tasks of little consequence, other times in decisions and processes with an important stake. The wide spectrum of uses have varying levels of impact and include everything from search engine ranking decisions, support to medical diagnosis, online advertising, investment decisions, recruitment decisions, autonomous vehicles and even autonomous weapons. This creates great opportunities but brings challenges that are amplified by the complexity of the topic and the relative lack of accessible research on the use and impact of algorithmic decision-making.

The aim of the **algo:aware** project is to provide an evidence-based assessment of the types of opportunities, problems and emerging issues raised by the use of algorithms in order to contribute to a wider, shared, and evidence-informed understanding of the role of algorithms in the context of online platforms. The study also aims to design or prototype policy solutions for a selection of issues identified.

The study was procured by the European Commission and is intended to inform EU policy-making, as well as build awareness with a wider audience.

The draft report should be seen as a starting point for discussion and is primarily based on desk-research and information gathered through participation in relevant events. In line with our methodology, this report is being published on the **algo:aware** website in order to gather views and opinions from a wide range of stakeholders on:

1) Are there any discussion points, challenges, initiatives etc. not included in this State-of-the-Art Report?
2) To what extent is the analysis contained within this report accurate and comprehensive? If not, why not?
3) To what extent do you agree with the prominence with which this report presents the various issues? Should certain topics receive greater or less focus?

Introduction

Algorithmic decision-making systems are deployed to enhance user experience, improve the quality of service provision and/or to maximise efficiencies in light of scarce resources in both public and commercial settings. Such instances include: a university using an algorithm to select prospective students; a fiscal authority detecting irregularities in tax declarations; a financial institution using algorithms to automatically detect fraudulent transactions; an internet service provider wishing to determine the optimal allocation of resources to serve its customers more effectively; or an oil company wishing to know from which wells it should extract oil in order to maximise profit. **Algorithms are thus fundamental enablers in modern society**.

The widespread application of algorithmic decision-making systems has been enabled by advancements in computing power and the increased ability to collect, store and utilise massive quantities and a variety of personal and non-personal data from both traditional and

non-traditional sources. Algorithmic systems are capable of integrating more sources of data, and identifying relationships between those data, more effectively than humans can. In particular, they may be able to detect rare outlier cases where humans cannot.

Moreover, algorithmic decision-making does not occur in a vacuum. It should be appreciated that qualifications regarding the types of input data and the circumstances where automated decision-making is applied are made by designers and commissioners (i.e. human actors). Given the emerging consensus that the use of algorithmic decision-making in both the public and private sectors is having, and will continue to have, profound social, economic, legal and political implications, civil society, researchers, policymakers and engaged industry players are debating whether the application of algorithmic decision-making is always appropriate.

Thus, **real tensions exist between the positive impacts and the risks presented by of algorithmic decision-making** in both current and future applications. In the European Union, a regulatory framework already governs some of these concerns. The General Data Protection Regulation establishes a set of rules governing the use of automated decision-making and profiling on the basis of personal data. Specific provisions are also included in the MiFID II regulation for high speed trading, and other emerging regulatory interventions are framing the use of algorithms in particular situations.

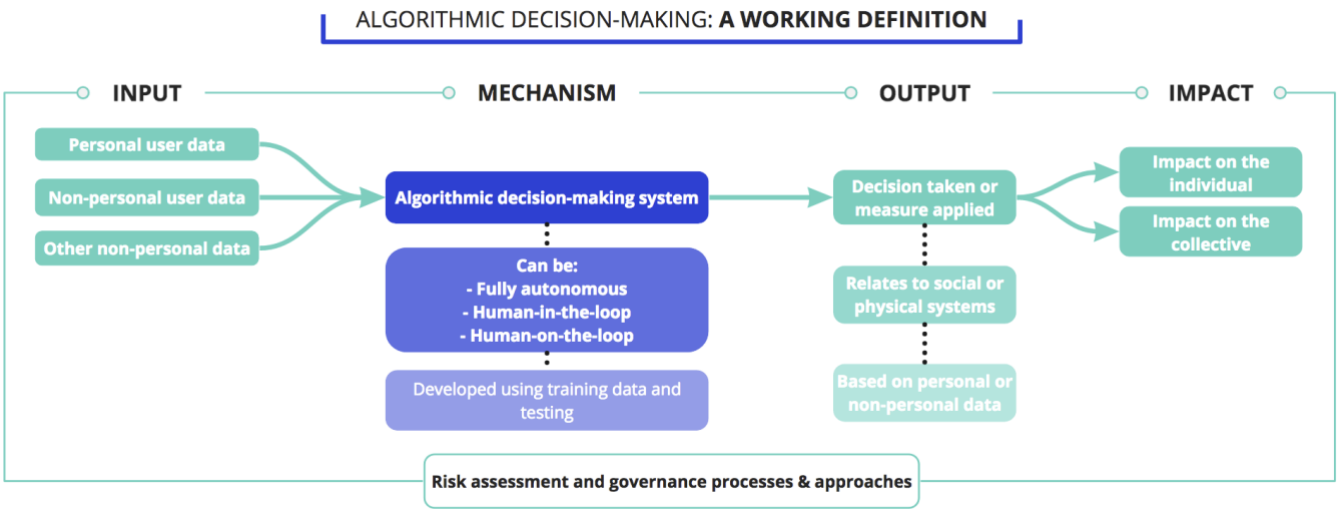
Scope of the report

The working definition for *decision-making algorithms*[1] in the scope of this report, and the outputs of **algo:aware** generally, is as follows:

> *A software system – including its testing, training and input data, as well as associated governance processes*[2] *– that, autonomously or with human involvement, takes decisions or applies measures relating to social or physical systems on the basis of personal or non-personal data, with impacts either at the individual or collective*[3] *level.*

The following figure represents the definition visually by mapping it to the parts of a 'model', typically comprising inputs, a processing component or mechanism and outputs.

[1] The definition of algorithmic decision-making is to be interpreted as a decision taken by a decision-making algorithm.
[2] Including risk and impact assessments, audit and bias histories, and associated risk management and governance processes.
[3] Such as impacts on financial markets and health systems, as well as impacts of algorithmic selection on online platforms.



Types of algorithms considered include, but are not limited to:

- Different types of search engines, including general, semantic, and meta search engines.
- Aggregation applications, such as news aggregators, which collect, categorise and re-group information from multiple sources into one single point of access
- Forecasting, profiling and recommendation applications, including targeted advertisements, selection of recommended products or content, personalised pricing and predictive policing
- Scoring applications (e.g. credit, news, social), including reputation-based systems, which gather and process feedback about the behaviour of users
- Content production applications (e.g. algorithmic journalism)
- Filtering and observation applications, such as spam filters, malware filters, and filters for detecting illegal content in online environments and platforms.
- Other 'sense-making' applications, crunching data and drawing insights.

The State-of-the-Art report analyses the academic literature and indexes a series of policy and regulatory initiatives, as well as industry and civil society-led projects and approaches.

Mapping the Academic Debate

There has been a wide array of academic engagement around the interaction of algorithmic systems and society.

Despite this, the concerns cited throughout the academic debate around algorithmic systems touch upon a huge array of areas of societal concern. Some of these are extensions of old challenges with added complexity from the changing and distributed nature of these technologies, such as liability concerns or societal discrimination. Others, however, seem newer, such as the transformation of mundane data into private or sensitive data, or the new and unusual ways in which technologies might fail or be compromised. Scholars from a wide variety of disciplines have weighed in on how these issues play out in a technical sense and how they see these issues in relation to governance, existing social and policy problems, societal framing and involvement in technological innovation, legal and regulatory frameworks and ethics. In many cases, these issues are not new, but they are reaching a level of salience and importance they did not previously hold.

The report structures the analysis along the following concepts, emerging as key concepts in the literature review and particularly useful to interrogate **whether the application of algorithmic decision-making systems bears societal risk and raises policy concerns**:

- **Fairness and equity** – in particular referring to the possible discriminatory results algorithmic decisions can lead to, and appropriate benchmarks automated systems should be assessed against;
- **Transparency and scrutiny** – algorithmic systems are complex and can make inferences based on large amounts of data where cause and effect are not intuitive. This concept relates to the potential oversight one might have on the systems;
- **Accountability** – a relational concept allowing stakeholders to interact, both to hold and to be held to account;
- **Robustness and resilience** – refers to the ability of an algorithmic system to continue operating the way it was intended to, in particular when re-purposed or re-used;
- **Privacy** – algorithmic systems can impact an individual's, or a group of individuals, right to private and family life and to the protection of their personal data; and
- **Liability** – questions of liability frequently arise in discussions about computational systems which have direct physical effects on the world (for instance self-driving cars).

Tensions exist between some of these concepts. Ensuring the transparency of an algorithmic system might come at the expense of its resilience, whilst ensuring fairness may necessitate a relinquishing a degree of privacy. Additional considerations on the role of the automated system and its performance compared to human-enabled decisions in similar applications give further contextualisation to the performance of algorithmic decision-making.

The main findings and outstanding questions identified in the literature are summarised as follows:

Fairness and equity. The literature has pointed to a number of instances where algorithmic decisions led to discriminatory results (e.g. against women in a given population), in particular due to inherent biases in historical data mirroring human bias. Fairness issues have a high profile in the academic literature, with a growing field of research and tools attempting to diagnose or mitigate the risks. Approaches range from procedural fairness concerning the input features, the decision process and the moral evaluation of the use of these features, to distributive fairness, with a focus on the outcomes of decision-making. Various approaches have also attempted to define a mathematical understanding of fairness in particular situations and based on given data sets, and to de-bias the algorithmic process through different methods, not without methodological challenges and trade-offs. In addition, a number of situations emerge which do not necessarily refer to decisions concerning specific individuals and unfair or illegal discrimination, but where different dimensions of fairness can be explored, possibly linked to market outcomes and impacts on market players, or behavioural nudging of individuals.

The report concludes on a series of emerging and remaining questions:

- What definitions of fairness are appropriate and necessary for different instances of algorithmic decisions? What are the tradeoffs between them? What are the fairness benchmarks for specific algorithmic decisions and in what situations should algorithms be held to a greater standard of fairness than human decisions? What governance can establish and enforce such standards? Do citizens and businesses feel that systems

which have been 'debiased' are more legitimate on the ground, and do such systems actually mitigate or reduce inequalities in practice?

Transparency and scrutiny. The comparative opacity of algorithmic systems has long led for calls for greater transparency from lawyers and computer scientists, and this has been reflected in both legislative developments and proposals across the world. The report presents several considerations as to the function and role of transparency in different cases and gives an overview of the controversy in the literature as to the different degrees of desired transparency for algorithmic systems compared to equivalent human decisions. It also discusses mitigating approaches, including development of simpler alternatives to complex algorithms, governance models including scrutiny, 'due process' set-up and oversight. It presents transparency models focusing on explainability approaches for complex models or disclosure of certain features, such as specific information on the performance of the model, information about the data set it builds on, and meaningful human oversight.

With a variety of approaches explored, questions emerge as to: What methods of transparency, particularly to society rather than just to individuals, might promote effective oversight over the growing number of algorithmic systems in use today?

Accountability is often undefined in the literature and used as an umbrella term for a variety of measures, including transparency, auditing and sanctions of algorithmic decision-makers. The report explores several models for accountability and raises a series of questions as to the appropriate governance models around different types of algorithmic decisions bearing different stakes.

Robustness and resilience. The academic literature flags several areas of potential vulnerability, stemming from the quality and provenance of data, re-use of algorithms or AI modules in contexts different than their initial development environment, or their use in different contexts, by different organisations, or, indeed, the unmanaged 'concept drift' where the deployment of the software does not keep up with the pattern change in the data flows feeding the algorithm. The robustness of algorithms is also challenged by 'adversarial' methods purposely studying the behaviour of the system and attempting to game the results, with different stakes and repercussions depending on the specific application area. Other concerns follow from attempts to extract and reconstruct a privately held model and expose trade secrets.

These areas are to a large extent underexplored and further research is needed. The **algo:aware** study will seek to further contextualise and details such concerns in analysing the specific case studies.

Privacy. A large part of the available literature focuses on privacy concerns, either to discuss and interpret the application of the General Data Protection Regulation, or to flag the regulatory vacuum in other jurisdictions. The report willingly de-emphasizes this corpus, arguably already brought to the public attention, and focuses on literature which addresses slightly different concerns around privacy. It flags emerging concerns around 'group privacy', closely related to group profiling algorithms, and flags possible vulnerabilities of 'leaking' personal data used to train algorithmic systems through attacks and attempts to invert models.

Liability. The report presents the different legal models of liability and responsibility around algorithmic systems, including strict liability, negligence-based liability, and alternative

reparatory policy approaches based on insurance schemes. It further explains situations where court cases have attributed liability for defamatory content on search engines.

Beyond this report, **algo:aware** will further explore some of these, and other questions that have been raised throughout this section, through sector/application-specific case studies. These case studies will subsequently form part of the evidence-base from which policy solutions may be designed. However, it seems unlikely that a single policy solution or approach will deal with all, or even most of those challenges currently identified. In order to address all of them, and to manage the trade-offs that arise, a layered variety of approaches are likely to be required. Civil society and industry have already begun to develop initiatives and design technical tools to address some the issues identified.

Initiatives from industry, civil society and other multi-disciplinary organisations

There is significant effort being directed towards tackling the challenges facing algorithmic decision-making by industry, civil society, academia and other interested parties. This is true across all categories of initiatives examined and relates to all of the perspectives discussed above. In particular, there are a large number of initiatives aimed at promoting responsible decision-making algorithms through codes of conduct, ethical principles or ethical frameworks.

Including this type of initiative, we have clustered the initiatives identified in four main types:

- **Standardisation efforts:** ISO and the IEEE are two of the most prominent global standards bodies, with the buy-in and cooperation of a significant number of national standards bodies. As such, it is important that these organisations are working towards tackling a number of these challenges. The final effort documented here, outside of the scope of the ISO and the IEEE, is the Chinese White Paper on Standardisation. Although no concrete work has been conducted, this document illustrates that stakeholders currently involved in the standardisation process in China – a multi-disciplinary group – are considering algorithmic decision-making from all the key perspectives being discussed.
- **Codes of conduct, ethical principles and frameworks:** As mentioned above, there are a vast number of attempts to govern the ethics of AI development and use with no clear understanding or reporting on take-up or impact. These initiatives have been initiated by stakeholders from all relevant groups, in some cases in isolation but also through multi-disciplinary efforts. Furthermore, much of this work attempts to tackle the challenges facing algorithmic decision-making from multiple perspectives. For instance, the ethical principles developed by the Software and Information Industry Association (SIIA) explicitly discuss the need for transparency and accountability; and the Asilomar Principles, which cover, in particular, topics of fairness, transparency, accountability, robustness and privacy. Interesting work that stands out and could be beneficial on a higher plane includes the work of Algorithmenethik on determining the success factors for a professional ethics code and the work of academics Cowls and Floridi, who recognised the emergence of numerous codes with similar principles and conducted an analysis across some of the most prominent examples. Cowls and Floridi's work is also valuable as it ties the industry of AI development and algorithmic decision-making to long established ethical principles from bioethics. The elements of learning these examples bring from established sectors can be extremely useful.

- **Working groups and committees:** The initiatives examined have primarily been initiated by civil society organisations (including, for example, AlgorithmWatch and the Machine Intelligence Research Institute) with the aim of bringing together a wide variety of stakeholders. Outputs of these initiatives tend to include collaborative events, such as the FAT/ML workshops, or research papers and advice, such as the World Wide Web Foundation's white paper series on *Opportunities and risks in emerging technologies*. As for the above, this type of initiative is often focused on tackling the challenges facing algorithmic decision-making from multiple perspectives. For instance, AlgorithmWatch maintains scientific working groups, which, in the context of various challenges, discuss, amongst others, topics of non-discrimination and bias, privacy and algorithmic robustness. Furthermore, no clear information on the impact of these initiatives is currently available.
- **Policy and technical tools:** In this category, the initiatives examined have been developed by academic research groups (e.g. the work of NYU's AI Now Institute and the UnBias research project), civil society (e.g. the Digital Decisions Tool of the Center for Democracy and Technology) or multi-disciplinary groups (e.g. the EthicsToolkit.ai developed through collaboration between academia and policy-makers). In terms of how these tools address the challenges facing algorithmic decision-making, they tend to focus on specific challenges; a clear example being the 'Fairness Toolkit', developed by the UnBias research project.

Policy initiatives and approaches

Across the globe, the majority of initiatives are very recent or still in development. Additionally, there are **limited concrete legislative or regulatory initiatives being implemented**. **This is not to say however that algorithmic decision-making operates in a deregulated environment**. The regulatory framework applied is generally technology-neutral, and rules applicable in specific sectors are not legally circumvented by the use of automated tools, as opposed to human decisions. Legal frameworks such as fundamental rights, national laws on non-discrimination, consumer protection legislation, competition law, safety standards still apply. Where concrete legislation has been enacted in the EU, the prominent examples relate primarily to the protection of personal data, primarily the EU's GDPR and national laws supporting the application of the Regulation. Jurisdictions such as the US have not yet implemented a comparable and comprehensive piece of legislation regulating personal rights. This might change to a certain extent with the introduction of the Future of AI bill, which includes more provisions on the appropriate use of algorithm-based decision-making. On the state level, the focus mainly is set on the prohibition of the use of non-disclosed AI bots (deriving from experiences of Russian AI bots intervening in the US Presidential election 2016) and the regulation of the use of automated decision-making by public administration.

The concept of algorithmic accountability should also be contextualized in the light of the policy initiatives. Indeed, the debate on accountability stems mainly from the United States, and while the societal aspects of the debate are very relevant and interesting, they reflect a situation where the legal context is very different than in the EU. The introduction of the GDPR means that a large part of the debate on accountability for processing of personal data is not as such relevant in the EU context. However, the practical application of the GDPR,

methodological concerns as to AI explainability, methods for risk and impact assessment, and practical governance questions are more pertinent to the EU debate.

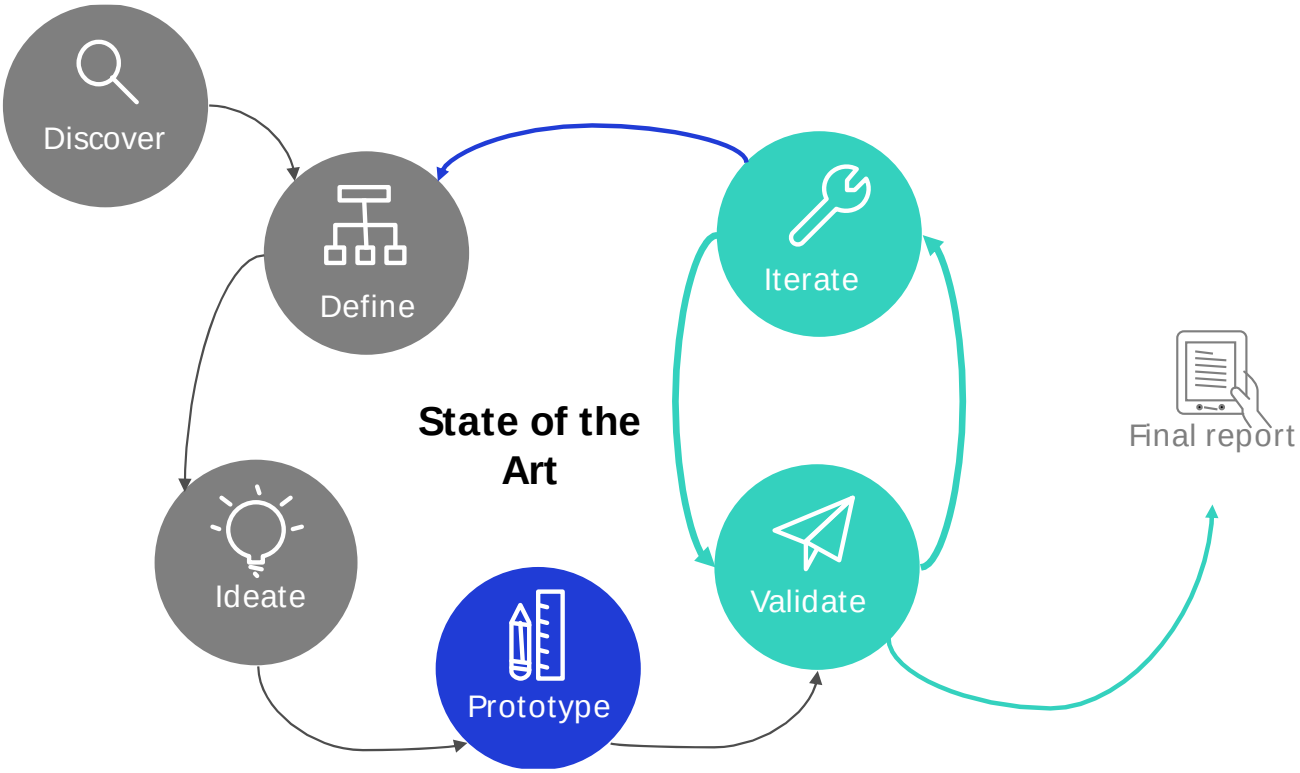
A few examples of AI-specific legislation have been identified, but the underlying question remains as to the need for assessing rule-making targeting a technology, or rather specific policy and regulatory environments adapted to the areas of application of the technology, and the consequent risks and stakes in each instance.

More commonly, however, the initiatives are softer in nature. These initiatives also reflect the aim of harnessing the potential of AI through the development of wide-reaching industrial and research strategies. Prominent types of initiatives implemented globally include:

- Development of **strategies on the use of AI and algorithmic decision-making**, with examples including France's *AI for Humanity* Strategy, which focuses on driving AI research, training and industry in France alongside the development of an ethical framework for AI to ensure, in particular, transparency, explainability and fairness. Another example is the Indian *National AI Strategy* and the EUR 3bn AI strategy issued by Germany in November 2018, which aims at making the country a frontrunner in the second AI wave, while maintaining strong ethical principals. Related to this are the numerous White Papers and reports developed, including the *German White Paper on AI*, the *Visegrád position paper on AI* and the Finnish *Age of AI* report.
- Establishment of **expert groups and guidance bodies** with examples including the Group of Experts and "Sages" established in Spain in 2018, the Italian *AI Task Force* and the German *Enquete Commission*. Considering the former example, this group has been tasked with guiding on the ethics of AI and Big Data through an examination of the social, juridicial and ethical implications of AI.

Next steps

This report represents an evolving account of the ongoing academic debate around the impacts of algorithmic decision-making, as well as a review of relevant initiatives within industry and civil society, and policy initiatives and approaches adopted by several EU and third countries. In line with the **algo:aware** design-led methodology, this version of the State-of-the-Art report should be considered the prototype. The purpose of the peer review methodology is to validate and provide inputs for the next iteration of the report.



Preamble

Algorithmic systems are changing all aspects of modern lives. This creates great opportunities and challenges which are amplified by the complexity of the topic and the relative lack of accessible research on the use and impact of algorithmic decision-making.

More and more decisions covering a wide spectrum of uses with varying levels of impact are being taken or supported by algorithms. These include search engine ranking decisions, medical diagnosis, online advertising, investment decisions or recruitment decisions, autonomous vehicles or even autonomous weapons.

The challenges of algorithmic decision-making have also captured public attention. The figure below provides a selection of headlines illustrating the way in which algorithms are being discussed in the media.

The aim of this study is to provide an evidence-based assessment of the types of opportunities, problems and emerging issues raised by the use of algorithms in order to contribute to a wider, shared understanding of the role of algorithms in the context of online platforms.

Finally, the study is intended to design or prototype policy solutions for a selection of issues identified.

The study was procured by the European Commission and is intended to inform EU policy-making as well as a wider audience.

This report presents the draft synthesis of the State-of-the-Art (SotA) in the field of algorithmic decision-making, focussing on the online environment and algorithmic selection/decision-making on online platforms. It presents the scope of the study, discusses definitions of

algorithmic decision-making and related concepts and provides insight into the academic debates on the topic, before illustrating the action being undertaken by civil society and industry and existing policy responses.

The draft report should be seen as a starting point for discussion and is primarily based on desk-research and information gathered through participation in relevant events. In line with our methodology, this report is being published in order to gather views and opinions from a wide range of stakeholders on the following questions:

4) **Are there any discussion points, challenges, initiatives etc. not included in this State-of-the-Art Report?**
5) **To what extent is the analysis contained within this report accurate and comprehensive? If not, why not?**
6) **To what extent do you agree with the prominence with which this report presents the various issues? Should certain topics receive greater or less focus?**

Our approach to gathering feedback comprises the following two overarching consultation streams:

- **Open crowd-sourced feedback:** Open to any and all interested parties, we are inviting any and all interested parties to submit feedback to the report via the **algo:aware** website.
- **Targeted peer-review consultation:** We will be conducting interviews with experts in the field, in particular including authors whose work is included in the bibliography.

In parallel, the information presented here will be tested, challenged and supplemented through workshops and additional participation of the research team at events. More information on the methodology for the peer-review process is available in this blog post: https://www.algoaware.eu/2018/12/04/sota-report-peer-review-methodology

We would like to encourage the readers to challenge our findings, provide us with additional information they might feel is missing and more generally engage with the study.

You can get in touch with the study team through the project's website: www.algoaware.eu or directly by e-mail: contact@algoaware.eu

1. Introduction and Context

Algorithms are an essential part of today's world and are applied in a wide range of processes and decision-making contexts. In many facets of daily modern living algorithmic decision-making systems have become pervasive and indeed fundamental. Certain industrial sectors have become dependent on their use such as the financial industry which utilises algorithms to automate trading decisions and detect investment opportunities. Indeed, algorithmic decision-making systems underpin economic growth in the digital economy and are already integrated in everyday technologies like smartphones that make predictions and determinations to facilitate the personalisation of experiences and advertisement of products.

Algorithmic decision-making systems are deployed to enhance user experience, improve the quality of service provision and/or to maximise efficiencies in light of scarce resources in both public and commercial settings. Such instances include: a university using an algorithm to select prospective students; a fiscal authority detecting irregularities in tax declarations; a financial institution using algorithms to automatically detect fraudulent transactions; an internet service provider wishing to determine the optimal allocation of resources to serve its customers more effectively; an oil company wishing to know from which wells it should extract oil in order to maximise profit. **Algorithms are thus fundamental enablers of numerous aspects in modern society**, contributing to increases in efficiency and effectiveness across all sectors of economic activity.

The widespread application of algorithmic decision-making systems has been enabled by advancements in computing power technology and the increased ability to collect, store and utilise massive quantities of personal and non-personal data from both traditional and non-traditional sources. Whilst this has presented citizens, businesses and governments with significant opportunities, it also has the capacity to have unintended negative consequences for individuals, vulnerable groups and trading dynamics. The speed at which the algorithmic decision-making technologies are being adopted, coupled with the scale of their potential impacts, naturally raises concerns, and even fears, as to the risks and mitigation of risks entailed by the take-up of the technology. However, some, if not most, algorithmic decisions are related to minute tasks of little consequence, whereas other instances raise policy and public attention.

The wide take-up of social media is a case in point, where the algorithms that underpin its functionalities have fundamentally changed the way citizens interact online, the way media is consumed, and the manner in which information is obtained from news outlets. Algorithms have been optimised by using historical patterns of engagement to predict and provide individuals with the most 'meaningful interactions', delivering posts, images, articles and advertisements that are deemed to be most relevant to the user. Such practices are argued by some to have led to the reinforcement of 'filter bubbles' which may influence the way citizens engage with democratic processes.

In online marketplaces, algorithms can influence or decide upon the manner in which products and services are recommended to users, as well as the order in which goods are ranked, depending on both search terms and historical user behaviour. This not only influences consumer choice, but also means businesses have an incentive to understand the types, and the weight attributed to input data if they wish to gain a competitive advantage in the increasingly popular online marketplaces. Indeed, the increased use of algorithmic decision-making systems in online environments is changing the organisational and operating models of businesses.

In the public sphere, take-up of automation and big data technologies shows a big potential for more effective policy-making and efficient service delivery at both national and local levels. At the same time, concerns related to the accuracy of the systems and risks of negative effects over the protection of fundamental rights also emerge, in particular in sensitive areas of application. For example, in the criminal justice system to predict the likelihood of recidivism, in policing to predict when and where there is an increased likelihood of crime being committed, or in risk assessments for interventions involving potentially vulnerable children.

Algorithms and algorithmic decision-making can **introduce consistency**, important not least for procedural purposes. Algorithmic systems are capable of integrating more sources of data, and identifying relationships between those data, more effectively than humans can. In particular, they may be able to detect rare outlier cases where humans cannot. These positive aspects of algorithms are also reflected in the list of 'Algorithm Pros' compiled by the DataEthics initiative[4], which includes the following summarised considerations:

- *Algorithms help humans make more rational decisions based on evidence and mathematically verified steps;*
- *Algorithms can aid governments in reducing bureaucratic monitoring and regulation;*
- *Algorithms aid individuals in managing their health in a more efficient way, creating less of a burden to national health systems. In this context, algorithms also hold great potential for advances in medicine and biomedical research, e.g. diabetes diagnostics, automation of surgical interventions.*

In support of these ideas Cowls and Floridi[5] highlight that underuse of AI due to fear, ignorance or underinvestment is likely to represent an opportunity cost, arguing that AI can foster human nature and its potentialities and thus create opportunities by enabling human self-realisation; enhancing human agency; increasing societal capabilities and cultivating societal cohesion.

Conversely, the authors conclude that good intentions gone awry, overuse or wilful misuse of AI technologies pose potential corresponding risks such as devaluing human skills; removing human responsibility; reducing human control and eroding human self-determination.

Algorithmic decision-making does not occur in a vacuum. It should be appreciated that qualifications regarding the types of input data and the circumstances where automated decision-making is applied, are made by designers and commissioners (i.e. human actors). Given the emerging consensus that the use of algorithmic decision-making in both the public and private sectors is having, and will continue to have profound social, economic, legal and political implications, civil society, researchers, policymakers and engaged industry players are debating whether the application of algorithmic decision-making is always appropriate.

There have been increased calls for **scrutiny** on the role that algorithms play where these determine information flows and influence public interest decisions that hitherto were exclusively handled by humans – especially in contexts that are of growing economic or societal importance. Algorithmic decision-making applied for surveillance/observation applications such as Raytheon's Rapid Information Overlay Technology (RIOT) gained prominence and was heavily criticised in the context of secret service surveillance. Scoring applications that gather and process feedback about participants' behaviour and derive ratings relating to such behaviour is being applied to sensitive areas such as credit scoring or social scoring. This has

[4] See https://dataethics.eu/en/prosconsai/
[5] See Cowls and Floridi (2018) Prolegomena to a White Paper on an Ethical Framework for a Good AI Society

also been criticised because of the considerable risks of social discrimination on the grounds of a person's race, age or religion, and further, may infringe personal privacy. Questions are emerging related to the governance and ethics of algorithmic decision-making, principles of fairness, reliability/accuracy and accountability, meaningful transparency, auditability and preserving privacy, freedom of expression or security.

Thus, **real tensions exist between the positive and negative impacts of algorithmic decision-making** in both current and future applications. In the European Union, a regulatory framework already governs some of these concerns. The General Data Protection Regulation establishes a set of rules governing the use of automated decision-making and profiling on the basis of personal data. Specific provisions are also included in the MiFID II regulation for high speed trading. The European Commission's Communication on Online Platforms further stated that greater transparency was needed for users to understand how the information presented to them is filtered, shaped or personalised, especially when this information forms the basis of purchasing decisions or influences their participation in civic or democratic life.[6] Transparency rules for ranking on online platforms are discussed in the context of the Commission's proposal for a Regulation on promoting fairness and transparency for business users of online intermediation services[7]. Other voluntary transparency provisions are included in the Code of Practice on disinformation[8], with a particular focus on political advertising.

The aim of this report is to synthesise the *State-of-the-Art* in the academic literature, policy initiatives and industry-led or civil society initiatives around algorithmic decision-making (with a particular focus on its application in online platforms). The inherent opportunities and value of algorithmic decision-making systems is undeniably attested by the global scale of adoption. While the study will focus in the next steps also on identifying and delineate specific opportunities, this report focuses more heavily on the challenges for the development and use of algorithmic systems, as identified in the literature.

The report is framed around the following concepts, which are useful to interrogate **whether the application of algorithmic decision-making systems bears societal risk and raises policy concerns[9]**:

- **Fairness and equity** – in particular referring to the discriminatory results algorithmic decisions can lead to;
- **Transparency and scrutiny** – algorithmic systems are complex and can make inferences based on large amounts of data where cause and effect are not intuitive; this concept relates to the potential oversight one might have on the systems;
- **Accountability** – a relational concept facilitating stakeholders to interact with decision-makers, to have them answer for their actions, and face consequences where appropriate;
- **Robustness and resilience** – refers to the ability of an algorithmic system to continue operating the way it was intended to, in particular when re-purposed or re-used;

[6] See communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on *online platforms and the Digital Single Market. Opportunities and Challenges for Europe* (2016): https://ec.europa.eu/digital-single-market/en/news/communication-online-platforms-and-digital-single-market-opportunities-and-challenges-europe

[7] https://ec.europa.eu/digital-single-market/en/news/regulation-promoting-fairness-and-transparency-business-users-online-intermediation-services

[8] https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[9] Each of these concepts is discussed in greater length in the section relating to the academic debate.

- **Privacy** - algorithmic systems can impact an individual's, or a group of individuals, right to private and family life and to the protection of their personal data; and
- **Liability** - questions of liability frequently arise in discussions about computational systems which have direct physical effects on the world (for instance self-driving cars).

As will be discussed later (see section 3), tensions exist between some of these concepts. Ensuring the transparency of an algorithmic system might come at the expense of its resilience, whilst ensuring fairness may necessitate a relinquishing a degree of privacy.

Acknowledging that the topic of algorithmic decision-making is placed at the confluence between a wide span of disciplines, this report aims to provide an updated account, in a succinct and accessible style, on current cross-sector issues brought about by algorithmic decision-making, proposed governance models and suggested solutions and approaches to emergent challenges. This report is thus structured as follows:

- A '**Scoping'** section which establish the analytical scope of this report by providing a working definition of algorithmic decision-making and other relevant terminology;
- An analytical literature review of the **academic debate** around the impacts of algorithmic decision-making with regard to the six concepts outlined above;
- An overview of different **initiatives in industry, civil society, and other multi-disciplinary organisations** looking to promote and ensure ethical design, assessment and deployment of algorithmic decision-making systems;
- An **index of current and potential policy approaches in the field**, in the geographical scope of the EU, selected Member States and selected third countries.

This report is to be considered as a living document with an accompanying dynamic bibliography, including a list of must-reads,[10] which will be iteratively updated in accordance with the evolution of the field[11].

[10] A dynamic bibliography containing the references provided in this report is available at https://www.algoaware.eu/bibliography/ and https://www.zotero.org/groups/2200076/algoaware/items/

[11] A methodology from how stakeholders will be consulted is provided here: https://www.algoaware.eu/2018/12/04/sota-report-peer-review-methodology

2. Scope

The aim of this section is to establish the analytical scope of this report by establishing a working definition of **algorithmic decision-making**, which will be consistent throughout this report as well as other **algo:aware** outputs. In doing so, this section highlights the conceptual differences between algorithmic decision-making and other terms which might be considered synonymous in other instances, such as **artificial intelligence** and **machine learning**.

The working definition of algorithmic decision-making provided in this section is not to be interpreted as a universal definition, but rather as a broad working definition which provides the fundamental basis for the analytical work presented throughout this report. In addition, 'decision' and 'decision-making' are meant here as broad terms that include, but are not limited to, sorting and filtering of information as well as selective information provision.

Definition of algorithmic decision-making

In the context of **algo:aware**, we consider that the main elements of a working algorithmic decision-making definition should:

- Not only account for software and code, but also acknowledge input, training and testing data as fundamental characteristics of decision-making by algorithms;
- Consider decisions made by algorithms on the basis of both personal and non-personal data;
- Be granular enough to allow for further characterisation as a 'fully autonomous' system, a 'human-in-the-loop' system, or a 'human-on-the-loop' system (defined below); that is, the definition should also include the type or level of human involvement in the decision-making process;
- Include governance processes, such as auditing and bias histories, risk and impact assessments, as well as risk management processes or approaches;
- Not only relate to the impacts of decision-making at the level of the individual, but also the impacts at the level of groups, markets, governments etc.;
- Refer to instances of human-machine and machine-machine interactions.

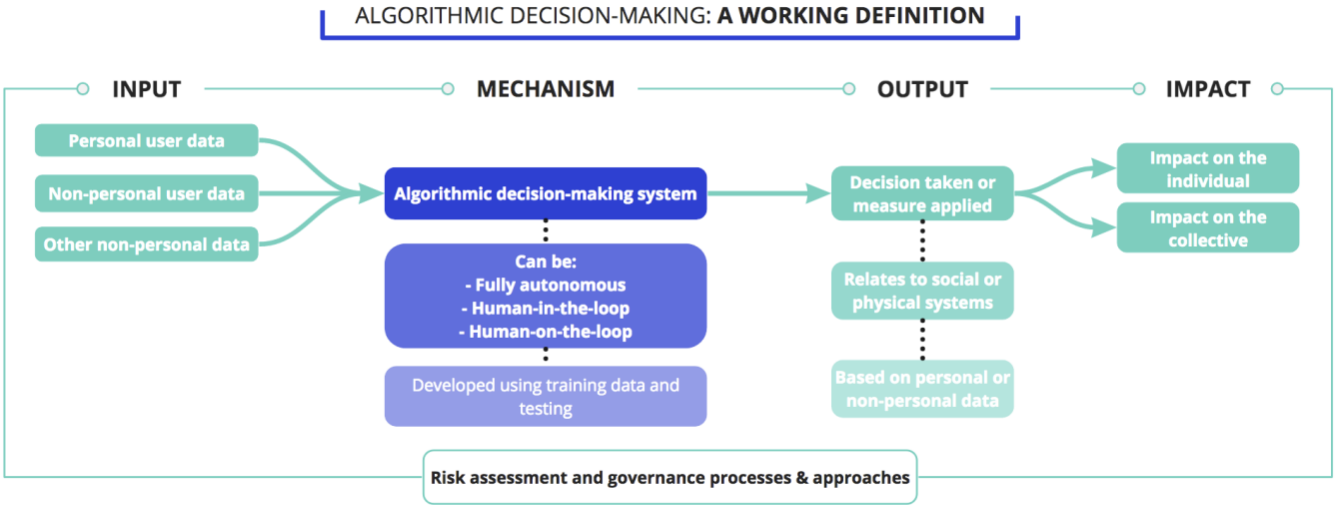
Thus, the working definition for *decision-making algorithms*[12] in the scope of this report, and all the outputs of **algo:aware**, is as follows:

A software system – including its testing, training and input data, as well as associated governance processes[13] – that, autonomously or with human involvement, takes decisions or applies measures relating to social or physical systems on the basis of personal and/or non-personal data, with impacts either at the individual or collective[14] level.

The following figure represents the definition visually by mapping it to the parts of a 'model', typically comprising inputs, a processing component or mechanism and outputs.

[12] The definition of algorithmic decision-making is to be interpreted as a decision taken by a decision-making algorithm.
[13] Including risk and impact assessments, audit and bias histories, and associated risk management and governance processes.
[14] Such as impacts on financial markets and health systems, as well as impacts of algorithmic selection on online platforms.

Figure 1: algo:aware working definition of algorithmic decision-making



Analytical scope

Building on the working definition, the remainder of this section establishes the analytical scope of the report by:

- Clarifying the **relationship between algorithmic decision-making and similar concepts**, such as 'Artificial Intelligence' (AI) and machine-learning. This is illustrated through the presentation of application examples.
- Expanding on **elements of the working definition** that benefit from further explanation, including: i) the **level of human involvement**; ii) the potential **negative impacts** of algorithmic decision-making; and iii) the **governance** of decision-making algorithms.

Algorithmic decision-making and other concepts

There are a range of technologies and concepts related to algorithmic decision-making, including in particular AI, machine learning, decision-support algorithms etc. This sub-section presents these concepts and illustrates how they relate to algorithmic decision-making.

'Artificial Intelligence' (AI), for instance, generally refers to **software-hardware systems that exhibit intelligent and meaningful behaviour in their context**, such as sensing and analysing their environment[15], having the ability to communicate, plan, learn and reason, as well as taking actions to achieve specific goals. AI has been the subject of scientific study since the 1950's. However, our faith in being able to harness AI as effectively as promised has fluctuated over the intervening years. This has resulted in periods of rapid progress and excitement, quickly followed by periods of decreased investment and interest – **periods known in the field as 'AI winters'**.[16]

[15] The term 'environment' is to be understood broadly in this instance, thus encompassing both virtual and physical environments.
[16] MIT Technology Review (2016) AI Winter isn't coming

Despite the ongoing debate on whether the potential economic and societal benefits of AI are over-publicised and exaggerated, potentially leading to another 'AI winter', the dominant opinion across industry and among policy leaders is that **AI is 'here to stay'**[17,18]. This view is evidenced by recent advances in the field, in particular in the following areas:

- **Perception:** a notable example is speech recognition. Speech recognition is now three times faster, on average, than typing on a mobile phone[19] and the average error rate for speech recognition has dropped from 8.5% to 4.9% since 2016[20].
- **Cognition:** advances in this area include: the development of systems capable of beating the best Go players in the world[21]; significant improvements in the cooling efficiency of data centres (by up to 15%)[22]; and improvements in AI-enabled money laundering detection systems[23].

The speed of improvement of AI has benefitted from the development of **machine learning**, including **deep learning** and **supervised learning**. Although machine learning can be used as a mechanism for achieving artificial intelligence, it is often treated as a separate field and many researchers apply machine learning to tackling problems of a practical nature with no need for intelligence.[24]

With that said, AI and machine learning systems can be based on a multiplicity of methods and algorithmic implementations. However, the majority of recent successes in this field falls into a particular class of systems, namely **supervised learning systems**, in which machines are given vast amounts of data in the form of correct examples of how to answer a particular problem[25], e.g. inputting images of various animals with correct output labels for each animal. The training data sets for these systems often consist of thousands or even millions of examples, each of which labelled with the correct answer.[26] Currently, the main driver of successful algorithmic implementations of these systems is **deep learning**. Deep learning algorithms have a great advantage over earlier versions of machine learning algorithms in that they can be trained and deployed on much larger data sets than their predecessors. In addition to supervised learning systems, the field of **reinforcement learning** has also recently grown in popularity. Reinforcement learning systems require a programmer to:

i) specify the current state of the system and its goal
ii) list allowable actions, and
iii) describe elements that constrain the outcomes of each action[27].

---

[17] MIT Technology Review (2016) AI Winter isn't coming
[18] European Commission (2018) 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, SWD(2018) 137 FINAL
[19] Ruan S. *et al.* (2017) Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones, *Journal Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1, 4
[20] Harvard Business Review | The Big Idea (2018) The Business of Artificial Intelligence.
[21] See https://www.bbc.co.uk/news/technology-35785875
[22] See https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/
[23] See https://www.technologyreview.com/s/545631/how-paypal-boosts-security-with-artificial-intelligence/
[24] Langley, P. (2011) The Changing Science of Machine Learning, Kluwer Academic Publishers. Last accessed on 29.11.2018 at: http://www.isle.org/~langley/papers/changes.mlj11.pdf.
[25] Harvard Business Review | The Big Idea (2018) The Business of Artificial Intelligence.
[26] Harvard Business Review | The Big Idea (2018) The Business of Artificial Intelligence.
[27] Harvard Business Review | The Big Idea (2018) The Business of Artificial Intelligence

The objective of this approach is for the system to learn how to achieve the specified objective by using a series of allowable actions. Reinforcement learning systems are thus particularly useful for instances in which a human programmer is able to specify a goal but not the path to achieving it.

Despite not being synonymous terms, we **consider AI, machine learning and algorithmic decision-making to be inter-related concepts**. In the context of the analysis presented in this report, we consider all algorithms associated with AI and machine learning properties (e.g. regression, classification, clustering, selection) to be decision-making algorithms. Moreover, we also consider that a subset of decision-making algorithms is not necessarily and explicitly 'intelligent', or AI-enabled, having their implicit decision-making criteria programmed in through relatively simple instructions.

The decision-making algorithms that underlie algorithmic selection processes, especially in online environments, are of particular interest to the analysis presented herein. This is because of the wide range of applications (and their technical, legal, and sectoral specificities) they concern, including[28],[29]:

- Different types of search engines, including general, semantic[30], and meta[31] search engines. In addition to general-purpose algorithmic search engines, there are a wide range of applications for special searching in particular domains or regarding particular issues (e.g. genealogy search engines), or search engines embedded on particular websites, market places, etc.
- Aggregation applications, such as news aggregators, which collect, categorise and re-group information from multiple sources into one single point of access[32,33,34];
- Forecasting, profiling and recommendation applications, including targeted advertisements, selection of recommended products or content, personalised pricing and predictive policing[35,36,37,38];

[28] Latzer *et al.* (2014) *The Economics of Algorithmic Selection on the Internet*, Institute of Mass Communication and Media Research Working papers, October 2014

[29] It is worth noting that the applications and systems referred in this report are made up of a diverse array of algorithms. These include, for instance, regression algorithms (e.g. linear, logistic and stepwise regression), decision tree algorithms (e.g. classification and regression tree), Bayesian algorithms (e.g. Naïve Bayes, Bayesian Belief Network), clustering algorithms (e.g. k-means, hierarchical clustering), and many others. For an extensive but non-exhaustive list, see https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

[30] Search engines that attempt to understand the user's intent and context in order to provide more relevant search results. H. Bast, B. Buchhold, E. Haussmann (2016) Semantic Search on Text and Knowledge Bases

[31] Search engines that use data from other search engines to produce search results. Giles et al. (1999), Architecture of a metasearch engine that supports user information needs.

[32] Zhu, H., M. D. Siegel and S. E. Madnick (2001), 'Information Aggregation: A Value-added E-Service'

[33] Aguila-Obra, A. R., A. Pandillo-Meléndez and C. Serarols-Tarrés (2007), 'Value creation and new intermediaries on Internet. An exploratory analysis of the online news industry and the web content aggregators', International Journal of Information Management, 27 (3), 187-199.

[34] Calin, M., C. Dellarocas, E. Palme and. J. Sutanto (2013), 'Attention Allocation in Information-Rich Environments: The Case of News Aggregtaors', Boston U. School of Management Research Paper No. 2013-4.

[35] Küsters, U., B. D. McCullough and M. Bell (2006), 'Forecasting software: Past, present and future', International Journal of Forecasting, 22 (3), 599-615

[36] Issenberg, S. (2012), 'The Victory Lab', New York: Crown Publishers.

[37] Silver, N. (2012), 'The Signal and the Noise: Why So Many Predictions Fail – but Some Don't', New York: Penguin

[38] Pathak, B. K., R. Garfinkel, R. D. Gopal, R. Venkathesan and F. Yin (2010), 'Empirical Analysis of the business value of Recommender Systems', Journal of Management Information Systems, 27 (2), 159-188.

- Scoring applications (e.g. credit, news, social), including reputation-based systems, which gather and process feedback about the behaviour of users, further deriving ratings and scores from this behavioural data;
- Content production applications (e.g. algorithmic journalism), that is, algorithmic systems that create content automatically[39,40,41,42];
- Filtering and observation applications, such as spam filters, malware filters, and filters for detecting illegal content in online environments and platforms. For instance, passive filters can select certain elements, but instead of displaying these to the user, they prevent access to them;
- Other 'sense-making' applications, crunching data and drawing insights.

Elements of a decision-making algorithm

Key elements of the working definition that require further explanation include: i) the **level of human involvement**; ii) the potential **negative impacts** of algorithmic decision-making; and iii) the **governance** of decision-making algorithms.

Regarding the **level of human involvement** in a decision-making algorithm, it is important to first clarify the line between algorithmic decision-making and algorithmic decision-support systems; a line that can be blurred depending on the level of human involvement. The difference between the two is presented in the below box.

Box 1: Key concepts – decision-support vs. decision-making algorithms

Key concepts: Decision-support vs. decision-making algorithms

Decision-support algorithms: do not take decisions in an automated way; they inform a human decision-maker.[43]

Decision-making algorithms: the output of the algorithmic operation itself results in a decision that is fully automated — a system that runs computationally also has the ability to trigger the action that the algorithm informs.

For algorithmic decision-making systems, OpenAI's Paul Christiano[44] details **three main types of decision-making algorithm based on the degree of human involvement** in the decision-making process[45]:

[39] See http://www.wired.com/2012/04/can-an-algorithm-write-abetter-news-story-than-a-human-reporter/
[40] Steiner, C. (2012), 'Automate This: How Algorithms Came to Rule Our World', New York (a.o.): Penguin.
[41] Anderson, C. W. (2013), 'Towards a Sociology of Computational and Algorithmic Journalism', New Media & Society, 15 (7), 1005-1021.
[42] Wallace, J. and K. Dörr (2015), 'Beyond Traditional Gatekeeping. How Algorithms and Users Restructure the Online Gatekeeping Process', Conference Paper, Digital Disruption to Journalism and Mass Communication Theory, 2-3 October 2014, Brussels.
[43] This does not mean that this human decision-maker is in a position to criticise the decision-support system meaningfully: automation bias (i.e. the over or under reliance on decision support) is an important psychological phenomenon linked to other biases.
[44] https://ai-alignment.com/counterfactual-human-in-the-loop-a7822e36f399
[45] Christiano also proposes a fourth type of autonomous system, the 'human-in-the-counterfactual-loop'. In this type of autonomous system, every time a decision is to be made the algorithm would flip a biased coin which would come up heads with a small probability (e.g. 0.001%). If the result is 'heads', the system consults

- **Fully autonomous systems**: operate without human supervision;
- **Human-in-the-loop systems**: exclusively follow specific human instructions;
- **Human-on-the-loop systems**: a human oversees the system and may override it.

Regardless of the level of human involvement, the **algorithmic decision-making applications listed above can carry significant negative impacts** at the level of individuals, vulnerable groups, markets and governments. For instance, filtering applications can be used to block political information in authoritarian regimes[46]; scoring applications can make profiling decisions that are discriminatory against minority groups or individuals or infringe personal privacy[47]; aggregation applications can have direct impacts on the profitability of media markets[48] (e.g. newspapers) and on intellectual property rights[49]. In relation to algorithmic selection in online environments, Michael Latzer and colleagues have mapped the specific social risks to eight categories:[50] i) manipulation; ii) diminishing variety, the creation of biases and distortions of reality; iii) constraints on the freedom of communication and expression; iv) threats to data protection and privacy; v) social discrimination; vi) violation of intellectual property rights; vii) possible transformations and adaptations of the human brain; and viii) uncertain effects of the power of algorithms on humans, e.g. growing independence of human control and growing human dependence on algorithms.

However, considering the full extent of algorithmic decision-making applications, **there is not a unique way to classify these impacts, risks and challenges.** As detailed through the following section, this report maps them to the concepts and current debates on fairness, accountability, transparency and scrutiny, robustness and resilience, privacy and liability.

As outlined above, the risks and challenges brought about by algorithmic decision-making systems and **applications often raise complex questions about governance in different contexts**. Algorithmic governance is meant here from two perspectives:

- 'Governance by algorithms' (i.e. the means through which the deployment of algorithms shapes varied aspects of society), and
- 'Governance of algorithms', which refers to the practices to control, shape and regulate algorithms[51].

---

a human, supplying them with context and recording feedback. Otherwise, the system does not consult with a human and attempts to make the decision the human would have made.

[46] Deibert, R., J. Palfrey, R. Rohozinski and J. Zittrain (eds) (2010), 'Access Controlled', Cambridge CA, London: MIT Press.

[47] Steinbrecher, S. (2006), 'Design Options for Privacy-Respecting Reputation Systems within Centralised Internet Communities', in S. Fischer-Hübner, K. Rannenberg, L. Yngström and S. Lindskog (eds), Security and Privacy in Dynamic Environments, Proceedings of the IFIP TC-11 21st International Information Security Conference (SEC 2006), 22–24 May 2006, Karlstad, Sweden: Springer, pp. 123-134.

[48] Weaver, A.B. (2013), 'Aggravated with Aggregators: Can International Copyright Law Help Save the Newsroom?', Emory International Law Review, 26 (4), 1159-1198.

[49] Isbell, K. (2010), 'The Rise of the News Aggregator: Legal Implications and Best Practices', Berkman Center for Internet & Society Research Publication 2010-10

[50] Latzer *et al.* (2014) *The Economics of Algorithmic Selection on the Internet*, Institute of Mass Communication and Media Research Working papers, October 2014

[51] Latzer et al. (2014) The Economics of Algorithmic Selection on the Internet, Institute of Mass Communication and Media Research Working papers, October 2014

The opportunities for a social shaping of algorithmic decision-making by means of governance have attracted increased attention in the academic research literature, particularly with regard to the governance of algorithmic selection search applications[52,53,54].

Moreover, various approaches to reduce risks and increase the benefits of algorithmic decision-making (and, in particular, algorithmic selection) have been identified, ranging from market mechanisms at one end, to command and control regulation by state authorities at the other[55]. The diversity and quantity of viable governance options proposed in the literature (e.g. self-organisation by individual companies; (collective) industry self-regulation; co-regulation – regulatory cooperation between state authorities and the industry) highlight that there are no one-size-fits-all solutions for the governance of algorithms[56]. In addition, they demonstrate that governance of algorithms (and by algorithms) goes beyond regulating (the design and implementation of) code and the technology itself and involves a wider evidence-based approach relying on risk and impact assessments, organisational approaches, and business models and strategies[57].

Several examples of governance in the context of algorithms have been developed recently. For instance, and as noted by Latzer *et al.* "disputes on certain practices and implications of news aggregation, search and algorithmic trading have resulted in regulatory provisions such as the German ancillary copyright law (BGBl. 2013, part 1, no. 23, p. 1161), the right to be forgotten for search engines in the EU (ECJ, judgment C-131/12 Google Spain vs. AEPD and Mario Costeja Gonzalez), and measures to prevent stock market crashes caused by algorithmic trading, e.g., the European Markets in Financial Instruments Directive (MiFID 2, 2014/65/EU)". [58]

Furthermore, several proposals for 'non-regulatory' governance approaches of algorithmic decision-making systems are provided in the literature. A relevant example is provided in the recent report published by the *AI NOW Institute*[59], which proposes a context-sensitive and 'governance-inclusive' approach to defining algorithmic decision-making systems, as well as exemplar definitions for specific contexts. In fact, the approach and definitions provided by the AI NOW Institute are centred on the development of Algorithmic Impact Assessments for use within governments and public agencies.

The definition for algorithmic decision-making presented and explained through this section establishes the analytical scope of this report; it also provides a comprehensive platform for analysing multi-disciplinary issues relating to the diverse impacts brought about by the design, deployment and governance of decision-making algorithms. Specifically, the definition considers the different approaches on these topics brought about by policy-making,

---

[52] Moffat, V. R. (2009), 'Regulating Search', Harvard Journal of Law & Technology, 22 (2), 475-513.
[53] Langford, A. (2013), 'Monopoly: Does Search Bias Warrant Antitrust or Regulatory Intervention?', Indiana Law Journal, 88 (4), 1559-1592.
[54] Lewandowski, D. (2014), 'Why We Need an Independent Index of the Web', in R. König and M. Rasch (eds), Society of the Query Reader: Reflections on Web Search, Amsterdam: Institute of Network Cultures.
[55] Latzer et al. (2014) The Economics of Algorithmic Selection on the Internet, Institute of Mass Communication and Media Research Working papers, October 2014
[56] Ibid.
[57] Brown, I. and C. Marsden, (2013), 'Regulating Code. Good Governance and Better Regulation in the Information Age.' Cambridge, M.A., London: MIT Press.
[58] Latzer et al. (2014) The Economics of Algorithmic Selection on the Internet, Institute of Mass Communication and Media Research Working papers, October 2014.
[59] Reisman et al. (AI NOW Institute) (2018), Algorithmic Impact Assessments: a practical framework for public accountability.

technological, governance and academic research perspectives. The definition, and consequently the analytical scope of the work presented herein, has the flexibility to include and go beyond the "impact of algorithms at the level of the individual". It considers impacts at the level of vulnerable groups, markets and governments, across different sectors and disciplines, through a holistic view that includes code, data, governance processes, and associated organisational structures and business models. This dynamic analytical scope thus allows for analytically "zooming in and out" when considering the cross-sectoral impacts of algorithmic decision-making from various perspectives.

With this said, the next section **provides an up to date account of the academic debate around the different impacts of algorithmic decision-making** from the perspectives of fairness and equity; transparency and scrutiny; accountability; robustness and resilience; privacy and liability.

3. The Academic Debate - an analytical literature review

There has been a wide array of academic engagement around issue of algorithmic systems and society. In this section, we outline a range of these debates. While there is no single way within which they can be categorised, we draw upon and group several issues in the subsections that follow.

Applied discussions of social challenges around algorithmic systems date back several decades. Anthropologists were early in explicitly considering how expert systems being deployed in contexts such as hospitals connected with individual and societal concerns,[60] linking to the excitement and promise of diagnostic support tools in the medical domain, as were ethicists and philosophers of computing.[61] More recently, computer scientists and lawyers joined the debate in both Europe[62] and in the US,[63] and the topic has become a heated and interdisciplinary area of concern and collaborative research.

In this section, we provide a narrative overview of several different and interwoven strands of literature touching upon algorithms and society from a number of perspectives: fairness and equity; transparency and scrutiny; accountability; robustness and resilience; privacy and liability. The section does not compare in detail 'performance' of algorithmic decisions and human decisions along these perspectives, but questions around the benchmarks of accuracy, fairness, accuracy and accountability will be important to address in the next steps of the project and the development of the policy toolbox.

3.1 Fairness and equity

Digital systems have long been subject to concerns that they might be discriminating unfairly against certain individuals and groups. In the early 2000s the term 'weblining' became used to refer to new forms of racist profiling through digital services.[64] Drawing upon 'redlining',[65] where entire neighbourhoods were determined as too 'high risk' for fundamental service provision, 'weblining' was largely used in relation to attempts to deny or reduce services provided to predominantly black neighbourhoods in the US, including in finance and e-

[60] Diana E Forsythe, 'Using Ethnography in the Design of an Explanation System' (1995) 8 Expert Systems with Applications 403; Diana E Forsythe, 'New Bottles, Old Wine: Hidden Cultural Assumptions in a Computerized Explanation System for Migraine Sufferers' (1996) 10 Medical Anthropology Quarterly 551.
[61] Helen Nissenbaum, 'Computing and Accountability' (1994) 37 Communications of the ACM 72
Helen Nissenbaum, 'Accountability in a Computerized Society' (1996) 2 Science and Engineering Ethics 25.
[62] Bart Custers (ed), Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases (Springer 2013); Mireille Hildebrandt and Serge Gutwirth (eds), Profiling the European Citizen: Cross-Disciplinary Perspectives (Springer 2008).
[63] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671.
[64] 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 at page 29.
[65] RE Dwyer (2015). Redlining. In The Blackwell Encyclopedia of Sociology, G. Ritzer (Ed.). doi:10.1002/9781405165518.wbeosr035.pub2

commerce sectors.[66] Today, fairness issues in algorithmic systems are a high profile issue, and a growing field of research and practice attempts to diagnose, mitigate and govern this area.[67]

Pursuing fairness of algorithmic models in many cases means necessarily paying a price in terms of the model's accuracy. Procedural fairness refers to the fairness of the decision-making processes (means) that lead to the outcomes. That is, the consideration of the input features used in the decision process, and evaluation of the moral judgments of humans regarding the use of these features[68]. However, much of the literature thus far has focused on distributive fairness which refers to the fairness of the outcomes (ends) of decision-making, due to the more tangible link to anti-discrimination laws.

In Europe, discrimination law distinguishes between 'direct' and 'indirect' discrimination. In the US, the same terms map onto 'disparate treatment' and 'disparate impact'.[69] Direct discrimination (or disparate treatment) occurs when discrimination is based on a protected characteristic, such as gender or ethnicity: that characteristic itself was the *basis* for a different decision. Indirect discrimination, or disparate impact, centres on the use of other data which may be correlated with certain protected attributes, and through its use, may disproportionately impact people sharing certain protected characteristics. In this case, other data act as a proxy for the protected characteristics. For example, web browsing history might proxy for gender, race or sexuality, without any of those categories having been explicitly declared.

More formally, and concerning race,[70] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin states (Article 2) that:

"(a) direct discrimination shall be taken to occur where one person is treated less favourably than another is, has been or would be treated in a comparable situation on grounds of racial or ethnic origin;

(b) indirect discrimination shall be taken to occur where an apparently neutral provision, criterion or practice would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary."

One of the primary focuses in the academic literature is the extent to which machine learning systems risk committing indirect discrimination or fostering disparate impact.[71] There are several ways they can do this.

---

[66] Marcia Stepanek, Weblining: Companies are using your personal data to limit your choices—and force you to pay more for products, Bloomberg Business Week, Apr. 3, 2000, at 26; Wells Fargo yanks "Community Calculator" service after ACORN lawsuit, Credit Union Times (July 19, 2000), https://perma.cc/XG79-9P74; Elliot Zaret & Brock N Meeks, Kozmo's digital dividing lines, MSNBC (Apr. 11, 2000); Kate Marquess, Redline may be going online, 86 ABA J. 8 at 81 (Aug. 2000).

[67] Rachel Courtland, 'Bias Detectives: The Researchers Striving to Make Algorithms Fair' (2018) 558 Nature 357 <http://www.nature.com/articles/d41586-018-05469-3> accessed 21 June 2018.

[68] Nina Grgic-Hlaca et al. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning (2018). Association for the Advancement of Artificial Intelligence

[69] Reuben Binns, 'Fairness in Machine Learning: Lessons from Political Philosophy', *Conference on Fairness, Accountability and Transparency (FAT* 2018)* (PMLR 2018).

[70] It is legal convention to use the term 'race' to describe ethnic groups, however it is usually noted, as here, that the use of such term does not condone theories that there is more than one human race.

[71] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671.

Machine learning systems, which power many algorithmic decision-making applications today, are trained primarily on historical data.[72] A common method to create such a system is to train it using historical decisions, such as job hiring decisions or loans, and try to get the system to best replicate the previous human decisions without requiring the expensive process they entail. The challenge here is that such historical data in itself is often rife with prejudice and bias. In the job sector, women may have been unfairly declined jobs, or offered lower salaries than men, and a system trained to think this was the 'correct' output will replicate undesirable patterns of the past.[73]

There are a range of reasons for the existence of this biased historical data. In some cases, the data might be based on a faithful reading or understanding of the rules of the phenomena at the time, but a reading that is nonetheless out of date today. Ethical standards and norms in society change. Many EU nations used to criminalise homosexuality to varying degrees. Such a decision in the justice sector may have been legally 'correct' to make at the time, but it is one that we ethically do not want to put into systems today. Similarly, discrimination on the basis of gender used to be permitted by the Gender Directive (2004/113/EC), until that provision was struck down in 2012 by the European Court of Justice.[74] Insurance providers would, as a result, have a challenge when using historical data before the entering into force of the *Test-Achats* decision (December 2012) in the training of machine learning systems. In other cases, decisions may have been based on conscious or unconscious bias on behalf of human-decision makers.

It may also be the case that there is some undesired truth in the historical data that has been collected. Crime rates, rates of defaulting on loans, rates of addiction or substance abuse, are not the same across all geographies or demographics. This is often a policy challenge which agencies attempt to mitigate. A system that more-or-less accurately reflects biases in the world that we wish to remove is potentially working against such policy efforts and locking individuals or communities into vicious circles that we do not wish to subject them to.

Past data is also not equally sampled. In many cases, data might exist to capture some aspect of a phenomenon rather than another, and this could have an undesirable effect. Systems that rely on data collected from expensive smartphones, for example, are only going to collect data from individuals and in areas where those devices are common.[75] Without careful consideration of this, it can be possible to ignore, omit, or over/under-represent whole geographic or demographic portions of society. Similarly, areas which have been historically over-policed or surveilled are likely to be areas which models are more certain about the outcomes of. At any applied uncertainty threshold, the increased certainty about these events might cause discriminatory outcomes. When subjective classifications are considered, such as what counts as offensive speech, and these classifications are made manually by human labellers, there are concerns that individual differences and biases may enter these systems and create unintended effects downstream.[76]

---

[72] Some systems in highly tamed and well-understood environments, such as those which play board and video games, or those which operate in physical environments, train against simulations, but given a technical inability to simulate the world, these have limited demonstrated use in complex social contexts.
[73] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671.
[74] Case C-236/09 Test-Achats [2012].
[75] Kate Crawford, 'The Hidden Biases in Big Data' [2013] Harvard Business Review <https://hbr.org/2013/04/the-hidden-biases-in-big-data> accessed 15 September 2018.
[76] Reuben Binns and others, 'Like trainer, like bot: Inheritance of bias in algorithmic content moderation' (2017) International Conference on Social Informatics (SocInfo 2017); Alex Rosenblat and others,

Categorisation and classification can also exacerbate issues of fairness and discrimination. For example, '[a]ggregating those who subscribe to different branches of a religious doctrine (e.g. Catholic, Protestant; Shia, Sunni) within a single overarching doctrine (Christian, Muslim) might collapse distinctions which are highly relevant to questions of fairness and discrimination within certain context'.[77] Such issues are far from new, and scholars critically examining science and technology have long been aware of the challenges of classifying individuals in a myriad of subject ways that may not represent them, or may exacerbate other harms and challenges.[78]

'Debiasing' approaches

Computer scientists have sought to deal with this is a number of ways, primarily by defining formal definitions of fairness, and attempting to modify machine learning processes to yield models to meet them.[79] A naive definition might be 'disparate impact' or 'statistical / demographic parity', which consider the overall percentage of positive/ negative classification rates between groups.[80] However, this is blunt, since it fails to account for discrimination which might be explainable in terms of apparently legitimate grounds. Attempting to enforce demographic parity between men and women in recidivism prediction systems, if men have higher reoffending rates, could result in women remaining in prison longer despite being less likely to reoffend, particularly if there were societally legitimate grounds.[81]

A range of more nuanced measures have been proposed, including; 'accuracy equity', which considers the overall accuracy of a predictive model for each group;[82] 'conditional accuracy equity', which considers the accuracy of a predictive model for each group, conditional on their predicted class; 'equality of opportunity', which considers whether individuals from each group are equally likely to achieve a desirable outcome if they meet the relevant threshold[83] and 'disparate mistreatment', a corollary which considers differences in false positive rates between groups.[84] Somewhat problematically, some of these intuitively plausible measures of fairness turn out to be mathematically impossible to satisfy simultaneously except in rare and contrived circumstances (ibid), and therefore hard choices between fairness metrics must be made before the technical work of detecting and mitigating unfairness can proceed. A legitimate question therefore, is whether it should be incumbent of predictive model developers to assess which notion of fairness is most appropriate for a given situation?

'Discriminating Tastes: Uber's Customer Ratings as Vehicles for Workplace Discrimination' (2017) 9 Policy & Internet 256.

[77] Michael Veale and Reuben Binns, 'Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data' (2017) 4 Big Data & Society 205395171774353.

[78] Geoffrey C Bowker and Susan Leigh Star. *Sorting things out: Classification and its consequences* (MIT Press 2000).

[79] For an accessible overview, see Solon Barocas, Moritz Hardt and Arvind Naranyan (2018) Fairness and Machine Learning (fairmlbook.org).

[80] Faisal Kamiran and Indre Zliobaite (2013) Explainable and Non-explainable Discrimination in Classification. In Discrmination and Privacy in the Information Society (Springer); Dino Pedereshi, Salvatore Ruggieri and Franco Turini (2008) Discrimination-aware data mining. KDD'08, ACM.

[81] Cynthia Dwork et al (2012) Fairness through awareness. ITCS'12.

[82] See e.g. Julia Angwin et al (2016) Machine bias. ProPublica.

[83] See Moritz Hardt et al (2016) Equality of Opportunity in Supervised Learning. NIPS'16.

[84] For an overview, see Alexandra Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments' (2017) 5(2) Big Data 153.

Having defined fairness, these proposals generally suggest that the next step is to 'de-bias' the model so that it optimises the fairness constraint, usually trading off against accuracy and other constraints (such as model complexity). This can be achieved by modifying either the underlying data used for training; modifying the learning process; or modifying the model after it has been learned (post-processing).[85]

As has been emphasised in the research literature, companies seeking to predict factors about individuals, such as eligibility for a service, using mundane-seeming data, might themselves unknowingly be using a largely invisible intermediate variable to make the decision[86]. These proxy variables may be sensitive or even illegal to be use, touching upon protected data such as ethnicity or political opinion, or politically sensitive data such as socioeconomic status. If controllers do not have access to 'true' sensitive data themselves, it becomes difficult for them to analyse whether systems are exhibiting bias or not. Consequently, while some are concerned that algorithmic systems are intentionally biased against certain groups (i.e. direct discrimination), a potentially more pervasive concern is that systems are deployed without appropriate scrutiny or oversight and have indirect discriminatory effects even if they did not have discriminatory intentions. Indeed, it is now appreciated that it is the governance framework around of the training and deployment algorithmic decision-making algorithmic systems that is crucial to limit bias and promote fairness.

Limits to 'debiasing'

Firstly, 'debiasing' methods can fast come into conflict with privacy aims. If systems are aimed to take into account and adjust for certain protected characteristics, it requires those sensitive data to be collected and retained.[87] Some emerging cryptographic methods, utilising new privacy-enhancing technologies such as *secure multiparty computation*, have been designed to overcome this, where individuals only give over encrypted versions of data such as their ethnicity, health status or sexuality, but it is still used in the debiasing methods described above.[88] Such methods have promise in certain sectors, but are still emerging and are yet to see deployment in practice.

While these 'debiasing' techniques do hold some promise in mitigating the potential biases of algorithmic decision-making systems, they also risk glossing over important issues of power and the underlying social processes of discrimination which give rise to questions of potential algorithmic bias in the first place. Ultimately, in order to know how much to correct for biases, model-builders need to understand the processes behind patterns in the data learned by the model. This may well require eliciting input from domain experts who are aware of the data collection processes and / or social context, and who can judge the extent to which de-biasing is needed to correct issues, whether they stem from biases in the data labelling process, or historical discrimination against the population in question.

---

[85] Faisal Kamian and others, 'Techniques for discrimination-free predictive models' In Discrimination and Privacy in The Information Society (Springer, B Custers ed.) 2013.
[86] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671.
[87] Michael Veale and Reuben Binns, 'Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data' (2017) 4 Big Data & Society 205395171774353.
[88] Niki Kilbertus and others, 'Blind Justice: Fairness with Encrypted Sensitive Attributes' (2018) ICML 2018. https://arxiv.org/abs/1806.03281

Furthermore, putting the onus of dealing with social problems on organisations deploying these systems (and the data scientists within them) risks depoliticising what are fundamentally political problems.

Fairness in the context of algorithmic decision-making has typically focused on unfair effects on individuals in their personal capacity, often through the lens of discrimination law. However, platforms also increasingly algorithmically structure the terms and means through which businesses engage in market activity, with potentially unfair effects. For instance, online intermediaries which aggregate and provide access to e.g. restaurants, or hotels could purposefully or inadvertently systematically favour certain businesses through the design of ranking algorithms. In some cases, such as price comparison websites, alongside common concerns that the platform may give greater prominence to businesses that pay higher commission, there may be other data- and algorithm-related forms of unfair advantage. To this end, the European Commission proposed a regulation on 'Fairness in platform-to-business relations', which imposes a series of transparency obligations, including disclosure of the main ranking parameters, on online intermediaries and search engines.[89]

Platforms can also alter prices for different consumers in ways some consumer feel is unfair. While price discrimination is a long-held area of study in economics, and there are arguments in favour of it from stances of efficiency, there is evidence that surreptitious discrimination creates unease and distrust among consumers.[90] At its extremes, microtargeting, based on an individual's inferred willingness to pay, might seem to border on extortion even if there is no clear correlation with a protected characteristic such as race or health status. Online price discrimination does however fall within the remit of data protection law in Europe insofar as it relies on the processing of personal data and therefore requires the identification of a lawful basis for data processing[91] and potentially, according to European data protection authorities, also for the ability for a user to contest it, if the price becomes prohibitively high as a result.[92]

A parallel debate to fairness and discrimination discussions, which is less explored in terms of policy options, surrounds the manipulative potential of machine learning systems. Some argue that organisations can use algorithmic systems in combination with behavioural change approaches such as 'nudging' to manipulate users in new ways. Such approaches have been described as 'hypernudging' by some scholars,[93] and public concerns around these areas recently came to a head in relation to alleged electoral manipulation by organisations such as Cambridge Analytica. Recently, the issue has been framed in terms of the externalities from optimisation systems in general, rather than fairness in a narrow sense, considering how outcomes that might be unfair could relate more deeply to the type of 'solution' chosen, and

[89] https://ec.europa.eu/info/law/better-regulation/initiatives/ares-2017-5222469_en

[90] Frederik Zuiderveen Borgesius and Joost Poort, 'Online Price Discrimination and EU Data Privacy Law' (2017) 40 Journal of Consumer Policy 347.

[91] Frederik Zuiderveen Borgesius, 'Personal Data Processing for Behavioural Targeting: Which Legal Basis?' (2015) 5 International Data Privacy Law 163.

[92] Michael Veale and Lilian Edwards, 'Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling' (2018) 34 Computer Law & Security Review 398 at page 401; Article 29 Data Protection Working Party (2018) Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.

[93] Karen Yeung, '"Hypernudge": Big Data as a Mode of Regulation by Design' (2017) 20 Information, Communication & Society 118

the logics of what is inside the system, what is outside, and what is being optimised upon, rather than considering discriminatory effects downstream.[94]

This section examines the notions of fairness and the considerations and limits to debiasing algorithmic decision-making systems in order to achieve fairer outcomes for individuals and businesses. The literature has tended to focus on the fairness of outcomes as this has a more tangible link to the anti-discrimination law already in place in the EU. The focus on debiasing is stemmed from the perceived risk that algorithmic decision-making systems can lead to indirect discrimination. Indeed, there are numerous examples of algorithmic decisions making systems leading to discriminatory outcomes. The anti-discriminatory legal regime is well established, so the question remains whether there are instances where discrimination caused algorithmically made decisions are not clearly covered in law. Further, are new policy approaches required to further mitigate against this risk of discrimination in instances that are unclear?

The section also highlights the consensus in the academic literature that trade-offs are to be made between the fairness and the accuracy of the algorithmic model, but also trade-offs between different *types* of fairness as it impossible to simultaneously to satisfy all kinds of fairness. The questions that emerge thus are what are the instances, if any, in which it would not be acceptable reduce accuracy in place of fairness? And further to whom should the algorithmically decision-making system be made fair to. It is argued that, at least in the public sector, policy preferences should be built into the system. For example, if mass-incarceration is a primary concern, in a society where there is an unfair disparity between the prison terms of two groups of people, reducing the prison terms of to that of the lower group may be a reasonable fairness goal.[95] Do certain types of deployers of algorithmic decision-making systems have a greater responsibility to develop fairer algorithms? Are there groups of citizens, or types of organisations, that there should be an emphasis to protect from unfair outcomes, and if, so under what circumstances (e.g. should there be an effort to ensure that AI does not entrench existing biases that other policy areas are attempting to mitigate against)? These are challenging questions, and it is unlikely and perhaps unrealistic that they can be solved by model developers. It has been suggested that these such matters of values and law can only be resolved by the political process.[96]

3.2 Transparency and scrutiny

Today's algorithmic systems have the potential to be significantly more complicated than traditional formalised decision-making systems. The comparative opacity these systems are attributed with has long led for calls for greater transparency from lawyers and computer scientists,[97] and this has been reflected in both legislative developments and proposals across the world.

---

[94] Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, Seda Gürses (2018) POTs: Protective Optimization Technologies. https://arxiv.org/abs/1806.02711

[95] Richard Berk et al. (2017) Fairness in Criminal Justice Risk Assessments: The State of the Art

[96] Ibid.

[97] Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008); Mireille Hildebrandt, 'The Dawn of a Critical Transparency Right for the Profiling Era' [2012] Digital Enlightenment Yearbook 2012 41.

Before systems themselves can be made transparent, it has to be considered whether individuals can tell that a decision or personalisation measure is taken algorithmically at all. Web technologies have long been possible to tailor algorithmically, developed in the field of adaptive hypermedia. Debates on the 'scrutability' of these adapting and learning online systems are old,[98] and given interest around micro-targeting and echo chambers or 'filter bubbles', today these debates are still very relevant. Many developers and user designers have been trained in the practice of 'seamless' design,[99] where algorithmic decisions should be part of a convenient and invisible process for the users. Today however, it is debated whether users' attention should instead be actively drawn to the 'seams' of their online environments — glimpses into the way the systems around them work, through which they can grasp and learn more about the way the world is being personalised.[100] In this case transparency would attempt to create awareness for the user about the way that they world is being shaped around them and for them. In the physical world too, where Internet of Things devices are increasingly ubiquitous and the personalisation possibilities there are also increasing, it can be difficult for an individual to know when data relating to them is being processed, or when their environment is being altered.[101] Search and rankings too prove difficult to demonstrate ordering and algorithmic effects in, as it can be challenging to show how a system would have been if its input data or logics were different. Given that no ranking is 'neutral', but always relies on rules which prioritised some entries over others, it is difficult to identify a suitable contrasting example to compare a ranking against. In this regard, transparency obligations of the proposed 'Fairness in platform-to-business relations' regulation are intended to enhance fairness for businesses that utilise online platform intermediaries. Another example relates to MiFID II where transparency obligations were put in place to mitigate against systemic risk and protect against market abuse. In this regard, increased transparency aims to enhance investor protection, reinforce confidence in the financial market, addresses previously unregulated areas, and ensure that supervisors are granted adequate powers to fulfil their duties.

Even when algorithmic systems are identified, there remains considerable academic debate over the standards to which we should hold decisions made or informed by them to. Some have argued that there is a need for much greater transparency of algorithmic systems, particularly due to the consequential effects they can have in society.[102] Firstly, there are split opinions over the quality and quantity of transparency that algorithmic systems deserve. Some 'worry that automated decision-making is being held to an unrealistically high [standard], possibly owing to an unrealistically high estimate of the degree of transparency attainable from human decision-makers'.[103] According to this strand of argument, explanations from algorithmic systems should mirror the depth and utility of explanations we would try to extract from humans in similar situations. Others argue that the application of established

[98] Judy Kay, 'Scrutable Adaptation: Because We Can and Must', *Adaptive Hypermedia and Adaptive Web-Based Systems* (Springer 2006).
[99] Marc Weiser, 'The World Is Not a Desktop' (1994) 1 Interactions 7.
[100] Motahhare Eslami and others, '"I Always Assumed That I Wasn't Really That close to [her]": Reasoning about Invisible Algorithms in News Feeds' ACM CHI 2015.
[101] Ewa Luger and Tom Rodden, 'An Informed View on Consent for UbiComp', ACM UbiComp '13. (2013); Lilian Edwards, 'Privacy, Security and Data Protection in Smart Cities' (2016) 2 European Data Protection Law Review 28.
[102] Frank Pasquale, 'Restoring Transparency to Automated Authority' (2011) 9 J. on Telecomm. & High Tech. L. 235.
[103] John Zerilli and others, 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?' [2018] Philosophy & Technology doi:10.1007/s13347-018-0330-6

transparency principles from areas such as administrative law would '[not be] a higher standard [...] but one that is adapted to the way that an algorithm-assisted decision is structured'.[104] In particular, as algorithms are external to users of decisions support systems, transparency mechanisms can help them build mental models needed to understand whether they are good, legal or ethical to use or not.[105] In this regard, heightened transparency would be necessary just to provide the same level of transparency over decisions and processes as legal systems currently provide. In this regard, the French Digital Republic Act 2016 (Loi Lemaire),[106] contains particular provisions on 'algorithmic treatment' of individuals by the public administration. The law, which pre-empted some provisions of the GDPR, places requirements on government bodies that make decisions solely or partially by algorithmic systems to give individuals certain transparency including the "treatment parameters and, where appropriate, their weighting, applied to the situation of the person concerned".

There has also been considerable debate as to whether more complex algorithmic systems are needed or useful. Some research has focussed on creating simpler systems with comparable utility to more complex ones.[107] Other, older research has focussed on extracting simpler rules from trained neural networks which might be used in practice in place of complex systems.[108] There has also been a general push-back against inference-based, machine learning models by some researchers, who claim that only models where the pathways of cause-and-effect are well-understood are capable of overcoming some of the tricky challenges that algorithmic systems are being applied to.[109]

Where complex systems are deployed, debates exist over whether explanation rights provide useful safeguards in practice. While they appear attractive, and have significant public support, some authors have expressed concern that they might provide a meaningless, non-actionable form of explanation that does little more to help deal with algorithmic harms than privacy policies individuals have little time to read.[110] Data and advanced analytics have been long thought to undermine the already-struggling traditional policy tools of consent in both the EU and elsewhere, as individuals have limited ability to understand the sensitive inferences that can arise from the collection of seemingly mundane data.[111] Even if data subjects are required to be told (as inferences are also considered personal data under data protection law), many

---

[104] Marion Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power' (2018) 376 Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 20170359.

[105] Max Van Kleek, William Seymour, Michael Veale, Reuben Binns, Nigel Shadbolt, 'The Need for Sensemaking in Networked Privacy and Algorithmic Responsibility'. ACM CHI 2018 Sensemaking Workshop, Montréal, Canada, 21–26 April 2018. http://discovery.ucl.ac.uk/10046886/

[106] Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.

[107] Berk Ustun and Cynthia Rudin, 'Supersparse Linear Integer Models for Optimized Medical Scoring Systems'; Jiaming Zeng, Berk Ustun and Cynthia Rudin, 'Interpretable Classification Models for Recidivism Prediction' (2017) 180 Journal of the Royal Statistical Society. Series A, 689.

[108] Alan B Tickle and others, 'The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks' (1998) 9 IEEE Transactions on Neural Networks 12; Robert Andrews, Joachim Diederich and Alan B Tickle, 'Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks' (12/1995) 8 Knowledge-Based Systems 373.

[109] Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books 2018).

[110] Lilian Edwards and Michael Veale (2017). Slave to the Algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. 16 Duke Law and Technology Review 18-84.

[111] Solon Barocas and Helen Nissenbaum, 'Big Data's End Run around Anonymity and Consent', *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (Cambridge University Press 2014).

data controllers themselves do not know clearly what sensitive data is being processed, particularly where it is being inferred as an intermediate step (see section above on Fairness), and thus are unable to provide effective transparency. Data controllers also have a poor record at facilitating transparency around data-driven systems using the rights that do exist,[112] and so the provision of these rights has been argued to be just as much of a problem as their exercise.

Evidence suggests that biases can shape the way information is presented. For example, online ranking systems on major job advertising platforms have demonstrated gender-based inequalities, mostly to the detriment of females[113]. It is important to note that bias does not emerge from an algorithm alone but also arises from the data that serves as the input as well as arising from the ranking system itself.

Researchers have proposed a framework for quantifying and illuminating the biases that may underpin search results in the context of political searches, enhancing transparency, and presents another example where the aim is to empower users who seek to adjust the search results.[114]

Ideas of collective explanation and scrutiny serve as a counterpoint to individual transparency. This is linked to a variety of debates, such as around ensuring 'due process' in algorithmic systems,[115] ensuring that collective and empowered oversight exists,[116] and giving workers and individuals greater say in the way systems are deployed.[117] Some researchers see promise in the collective use of individual rights to achieve greater societal transparency.[118]

Researchers have attempted to design different explanation systems, also called 'explanation facilities', to deal with opaque machines. Some of the earliest approaches, particularly in the time of expert systems (which tried to put the expertise of say lawyers or doctors into question-answer machines) sought to extract rules or logics from neural networks and other machine learning models.[119] These were often used to help understand a domain better using data mining, although mixed approaches with understandable rules alongside more opaque neural

[112] Jef Ausloos and Pierre Dewitte, 'Shattering One-Way Mirrors – Data Subject Access Rights in Practice' (2018) 8 International Data Privacy Law 4 <https://academic.oup.com/idpl/article/8/1/4/4922871> accessed 3 May 2018.
[113] Le Chen *et al.* (2018). Investigating the Impact of Gender on Rank in Resume Search Engines. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems
[114] Juhi Kulshrestha *et al.* (2017). Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. CSCW '17 Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing
[115] Kate Crawford and Jason Schultz, 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms' [2014] BCL Rev; Danielle Keats Citron, 'Technological Due Process' (2008) 85 Washington University Law Review 1249.
[116] Lilian Edwards and Michael Veale, 'Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?' (2018) 16 IEEE Security & Privacy 46.<https://ssrn.com/abstract=3052831>; Jakko Kemper and Daan Kolkman, 'Transparent to Whom? No Algorithmic Accountability without a Critical Audience' [2018] Information, Communication and Society 1.
[117] Dan McQuillan, 'People's Councils for Ethical Machine Learning' (2018) 4 Social Media + Society 2056305118768303.
[118] René LP Mahieu, Hadi Asghari and Michel van Eeten, 'Collectively Exercising the Right of Access: Individual Effort, Societal Effect' (2018) 7 Internet Policy Review.
[119] Alan B Tickle and others, 'The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks' (1998) 9 IEEE Transactions on Neural Networks 12.

networks have an older history in AI and law.[120] These explanation facilities were not aimed at those affected by a system in general, but targeted at the users of the system.[121] In particular, such user-facing explanations can be used in practice to get buy-in or trust from these users by assuring them the system logic is something that they understand and can empathise with.[122] In these instances transparency is serving to enhance stakeholder trust in the algorithmic decision-making system.

Yet as systems have become more complex, particularly when dealing with complex areas such as images, where every pixel represents a data point that can take many colours, and there are many different image classification possibilities, it has become clear that it is futile to try to 'explain' the entire system in one go. Instead of trying to explain this model or explain the process by which it was built—a 'model-centric explanation'—many explanations have focussed on individuals and individual records—'subject-centric explanations'.[123]

A wide range of approaches have which tried to highlight what factors led to a particular decision. Some of these deal with more continuous domains, such as highlighting parts of the data which are the most 'important' to the classification.[124] There is also a history of explaining computer systems by which factors would need to be different in order for the system to have classified a case in a different way. These 'why not',[125] 'sensitivity'[126] or 'counterfactual'[127] explanations have promise in simple areas, but also bring concerns that their simplicity might allow them to be gamed, that they might produce impossible results or ask for impossible changes,[128] or that they will not be effective when the variables in the model themselves do not have human interpretable meanings. A wide variety of other explanation facilities have been developed, all of which are suited in different ways to different domains.[129] Recommender

---

[120] John Zeleznikow and Andrew Stranieri, 'The Split-up System: Integrating Neural Networks and Rule-Based Reasoning in the Legal Domain' (ACM Press 1995); J Zeleznikow, 'The Split-up Project: Induction, Context and Knowledge Discovery in Law' (2004) 3 Law, Probability and Risk 147.

[121] 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 <https://osf.io/97upg>.

[122] Michael Veale, Max Van Kleek and Reuben Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making' CHI 2018 (ACM Press 2018) https://doi.org/10.1145/3173574.3174014

[123] 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 <https://osf.io/97upg>.

[124] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, 'Model-Agnostic Interpretability of Machine Learning' [2016] arXiv:1606.05386 [cs, stat] <http://arxiv.org/abs/1606.05386> accessed 9 August 2018; Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, 'Why Should I Trust You?: Explaining the Predictions of Any Classifier' *KDD'2016* (ACM 2016); Grégoire Montavon and others, 'Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition' (2017) 65 Pattern Recognition 211.

[125] B Y Lim and AK Dey. Assessing demand for intelligibility in context-aware applications. UbiComp 2009; Lim B Y Lim, A K Dey, D Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. CHI 2009.

[126] 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. https://osf.io/97upg; Reuben Binns and others, '"It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions' CHI 2018 https://doi.org/10.1145/3173574.3173951.

[127] S Wachter, B Mittelstadt C Russell (2018) Counterfactual explanations without opening the black box. HJOLT.

[128] Berk Ustun, Alexander Spangher, Yang Liu (2019) Actionable Recourse in Linear Classification. Proceedings of ACM FAT* 2019. https://arxiv.org/abs/1809.06514

[129] See e.g. A Datta, S Sen and Y Zick, 'Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems', *2016 IEEE Symposium on Security and Privacy (SP)* (2016) <http://dx.doi.org/10.1109/SP.2016.42>

systems in particular have a long history of building explanation systems aimed at consumers,[130] while a similarly rich history exists in many other fields.[131]

Users have different perceptions of machine learning explanations, and these are only just beginning to be explored. Given that machine learning systems try to look at historical data and predict the future, it is reasonable to think that a type of explanation of the form 'you were predicted X, because the following cases similar to you had the characteristic X'—a 'case-based explanation'—would be reasonable. Yet individuals in practice appear to find such explanations unwanted and dislike them, perhaps because they do not feel they are being treated or valued as an individual.[132] As a result, testing explanations with real users is important, but something that is only just beginning to be done. Some research is beginning to indicate that involving users in the designs of explanations and making sure they relate to personal characteristics they identify with, might be better than proposing 'simple' explanations that seem compelling on paper.[133]

Most of these systems focus on the model of algorithm itself. In recent years, debate has been turning to the context that the model was built in, in addition to the structure and software behind the algorithmic system.[134] This might include information on how the model performs, what safeguards or tests it went through, how often it is retrained, on what dataset it was trained on, the oversight processes, and more. In particular, when systems are used as decision-support, there have been calls to ensure that any human input it meaningful[135] in order to avoid automation bias, where individuals under- or over-rely on automated systems.[136] The effect of greater transparency in this case may to inform the appropriate stakeholders perhaps in the context of liability.

Several outstanding issues emerge from the literature explored in this section. For example, what are the appropriate standards of de-basing appropriate for specific types of algorithmic decisions? What models of risk management and governance can mitigate and correct biases? Is there greater promise in collective explanation rights to algorithmic transparency in addition to the level of the individual? What are effective mechanisms of explaining algorithmically made decisions to individuals – 'model-centric', 'subject-centric', 'case-based' explanations for example? What models to inspect or audit of decision-making algorithms can serve to enhance trust and transparency in decision-making systems? What mechanisms can be deployed to

---

[130] Nava Tintarev, 'Explaining Recommendations' in Cristina Conati, Kathleen McCoy and Georgios Paliouras (eds), *User Modeling 2007*, vol 4511 (Springer Berlin Heidelberg 2007); Nava Tintarev and Judith Masthoff, 'Explaining Recommendations: Design and Evaluation' in Francesco Ricci, Lior Rokach and Bracha Shapira (eds), *Recommender Systems Handbook* (Springer 2015).

[131] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, Fosca Giannotti (2018). A Survey Of Methods For Explaining Black Box Models https://arxiv.org/abs/1802.01933

[132] Reuben Binns and others, '"It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions' CHI 2018 https://doi.org/10.1145/3173574.3173951

[133] Motohare Eslami et al (2018) 'Communicating Algorithmic Process in Online Behavioral Advertising' CHI 2018.

[134] Andrew Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' [2018] draft available on SSRN; Edwards and Veale (n 75).

[135] Michael Veale and Lilian Edwards, 'Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling' (2018) 34 Computer Law & Security Review 398.

[136] Kate Goddard, Abdul Roudsari and Jeremy C Wyatt, 'Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators' (01/2012) 19 Journal of the American Medical Informatics Association: JAMIA 121; Linda J Skitka, Kathleen L Mosier and Mark Burdick, 'Does Automation Bias Decision-Making?' (1999) 51 International Journal of Human-Computer Studies 991.

minimise automation bias, especially if algorithmic systems are deployed in situations where the human-in- the-loop is a non-specialist?

3.3 Accountability

Accountability can be defined in varying ways. In the literature on 'fairness, accountability, and transparency in machine learning', it is arguably the least discussed and defined term. While the term is frequently referred to in the context of algorithmic systems, it is often undefined and used as an umbrella term for a variety of measures, including transparency, auditing and sanctions of algorithmic decision-makers.

Some clarity may be achieved if we look to work on accountability from the public administration literature. One such work defines accountability as 'a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences'.[137] This definition is a *relational one*; it does not describe a feature of a system, but rather the existence of certain arrangements between various actors. Thus, making algorithmic systems accountable is less a question of building them in the right way, and more about putting in place certain institutional practices, regulatory frameworks and opportunities for stakeholders to interact, both to hold and to be held to account.[138]

Other approaches to accountability attempt to define it in a more black-and-white manner. In particular, some researchers are concerned that models will be 'swapped' without notice. In particular, those that are audited (e.g. for anti-discrimination) may not be the same as those used in practice.[139] Different approaches have been taken to approach this issue. Some have proposed that algorithmic systems be designed in such a way that they can be cryptographically checked to ensure that they are as they seem, through methods which differ depending on the system in question.[140]

Relatedly, recent work in machine learning and security attempts to provide frameworks for 'model governance', defined by Sridhar *et al.* as the "ability to determine the creation path, subsequent usage, and consequent outcomes of an ML model".[141] Similar work on 'decision provenance' borrows concepts from semantic web technologies, to track the inputs and effects of decisions taken within an algorithmic decision-making system.[142] Provenance information has a long standing history in supporting and enhancing the rigour of scientific research. However, the potential role of provenance information in facilitating algorithm transparency and accountability is understood to a lesser extent.

---

[137] Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework1' (2007) 13 European Law Journal 447.

[138] How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence (2017). *CNIL*

[139] Joshua A Kroll and others, 'Accountable Algorithms' (2016) 165 University of Pennsylvania Law Review.

[140] Joshua A Kroll, 'Accountable Algorithms' (PhD, Princeton University 2015); Niki Kilbertus and others, 'Blind Justice: Fairness with Encrypted Sensitive Attributes' (2018) ICML 2018. https://arxiv.org/abs/1806.03281

[141] Sridhar, Vinay, et al. "Model Governance: Reducing the Anarchy of Production ML." 2018 USENIX Annual Technical Conference (USENIX ATC 18). USENIX Association, 2018.

[142] Singh, Jatinder, Jennifer Cobbe, and Chris Norval. "Decision Provenance: Capturing data flow for accountable systems." arXiv preprint arXiv:1804.05741 (2018).

In these approaches, the common steps in development of a system are tracked and represented in terms of data pipelines and policy actions, such that any decision made by the system can be traced back to a set of design and policy choices. By capturing such provenance information, these approaches aim to raise levels of accountability.

Taken together there appear to be opportunities enhance accountability in interacting and complex algorithmic decision-making systems. It is suggested that it is important to establish which actors within the system who would benefit from accountability; a supervising body with legal authority to issue sanctions and/or impacted individuals of the system? And further, under which circumstances is the presently a gap in accountability? What role might provenance technologies play in supporting these broader goals? How can such solutions may go some way to answering how we can ensure that the data chosen is used responsibly and diligently? Further, what mechanisms can be provided to allow end users and developers to understand the algorithmic impact of their data choices?

3.4 Robustness and resilience

Machine learning and other algorithmic decision-making systems are no longer always deployed in the same place they are made. They are commonly packaged and transferred from academic research projects and in-house projects of large organisations. One driver of this is that it is legally easier to move a model rather than a personal dataset, particularly in the context of the strengthening of data protection law. Data may be collected from multiple different contexts by several organisations, re-purposed by others to train a system, which is then purchased or subscribed to by third parties or the public, either through licensing of APIs or trading of packaged models. This diversification of the pipeline opens up new possibilities for nefarious mis-use of algorithmic systems by outside actors.

Machine learning algorithms can fail in odd and unexpected ways, proving challenging for algorithm awareness. One recently-publicised form of attack is that of 'adversarial examples' — instances designed to fool a system into mis-classifying or mis-predicting them — such as an image of a turtle which is consistently mis-classified as a gun by an otherwise highly accurate image classifier.[143] Some of these attacks can be designed to fool algorithms in a specific way to achieve a desired outcome, such as convincing a detection system that offensive speech is in fact inoffensive, and should not be filtered out. Such natural language processing systems take strings of text and analyse them to try and classify them into particular groups. In practice, this might be as simple as removing spaces between words and adding 'love' onto the end, which can fool a system into thinking that a sentence is inoffensive, even though it remains easy to read and interpret by a human reader in its original form. This can be difficult to mitigate against in practice.[144] This has also been shown in the physical world. Specially printed multicoloured glasses, designed using a computational method, can fool can neural-network-powered image recognition systems not into just not recognising someone, but recognising

[143] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016

[144] Tommi Gröndahl and others, *All You Need Is 'Love': Evading Hate-Speech Detection* (28 August 2018) <http://arxiv.org/abs/1808.09115>.

them with very high certainty as a specific, different individual.[145] In other cases, it is possible to simply make the system very uncertain about the output and classify it in random and unpredictable ways, such as by adding particular types of noise.[146]

Other dangers include model extraction, which allows an attacker to reconstruct a private model, with only external access to the outputs of the model, rather than to the underlying model itself; and model inversion and membership inference attacks, which allow the attacker to learn otherwise private information about the individuals whose data was used to train the model. A key question is then whether the algorithms themselves might be more appropriately seen as personal data with some de-identification safeguards, rather than just as a software system or trade secret.[147]

In addition to these security-oriented concerns, there are numerous safety and reliability problems associated with current machine learning deployment. The notion of 'technical debt' describes the way that system designers may choose to postpone certain activities - such as organising and documenting their code in more accessible ways - which ultimately lead to higher maintenance costs further down the line. This is particularly tempting in projects involving machine learning; while it offers a "powerful toolkit for building complex systems quickly… it is remarkably easy to incur massive ongoing maintenance costs at the system level …". Risks include "boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, changes in the external world, and a variety of system-level anti-patterns."[148] The result is that many systems which initially work as intended at reasonable costs may quickly become unmanageable and costly. Algorithmic systems can reduce costs for businesses and other organisations by allowing existing tasks to scale in new ways, but they also bring new costs and the need for new roles in managing the way these systems interact with the environment.

Many of these issues of technical debt encompass how the algorithmic system's environment can change over time. Research in the public sector uses of machine learning has illustrated that such debt is common, particularly in the context of changing streams of data.[149] In technical terms, this is known as 'concept drift', which refers to when the 'concepts' the algorithm is trying to encode in the world change (such as societal preferences), can be hard to detect and react to.[150] It is not always clear when an algorithmic system is learning more about the world: such as being exposed to new customers, preferences or patterns that always

---

[145] Mahmood Sharif and others, 'Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition', *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016).

[146] Kenneth T Co, Luis Muñoz-González and Emil C Lupu, *Procedural Noise Adversarial Examples for Black-Box Attacks on Deep Neural Networks* (2018) <http://arxiv.org/abs/1810.00470>.

[147] For model theft e.g. Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. 2016 Stealing Machine Learning Models via Prediction APIs. In USENIX Security Symposium, pp. 601–618. For analysis of the implications for data protection, see Michael Veale, Reuben Binns, and Lilian Edwards, 'Algorithms that Remember: Model Inversion Attacks and Data Protection Law' (2018) 376 Philosophical Transactions of the Royal Society A 20180083 http://doi.org/10.1098/rsta.2018.0083.

[148] D Sculley and others, 'Hidden technical debt in machine learning systems' (2015) NIPS'15.

[149] Michael Veale, Max Van Kleek and Reuben Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making' CHI 2018 (ACM Press 2018) https://doi.org/10.1145/3173574.3174014

[150] J Gama and others, 'A Survey on Concept Drift Adaptation' (2013) 1 ACM Computing Surveys; Indrė Žliobaitė, Mykola Pechenizkiy and João Gama, 'An Overview of Concept Drift Applications' in Nathalie Japkowicz and Jerzy Stefanowski (eds), *Big Data Analysis: New Algorithms for a New Society* (Springer 2016).

existed but it had not incorporated as training data yet, or when it is seeing permanent change that means it should effectively 'forget' much of its past experience. As laws, norms and preferences change, it can be hard for machine learning systems, which do not understand broader context, to keep up.

Such questions of security, robustness and resilience are likely to compound the previous concerns of fairness, transparency and accountability, simply by adding in additional constraints which may be hard to manage in complex, fast-changing environments. Some practitioners have concerns that transparency can undermine resilience for example, by allowing end-users and decision-support users to better 'game' the system,[151] or by allowing models to be 'stolen' or 'reconstructed' using explanation facilities.[152] The use of algorithmic decision-making systems in society is for the most part predicated on the belief by both decision makers and decision subjects that they are reliable. If the issues related to robustness and resilience as highlighted above become common place in may undermine the level of trust required for their continued use. This is exemplified in the online advertising industry where the practice of algorithmically trading advertising space in an open marketplace has declined in recent years due to the prevalence of advertising fraud. It raises the question of whether a collective undertaking from academia and industry is necessary to mitigate against some of these emerging risks, and if so, what would this look like in practice? In general, however, such trade-offs are not completely understood, and still an area of highly active study.

3.5 Privacy

Algorithmic systems have the potential to transform seemingly non-sensitive data into sensitive data about individuals. At times, as discussed above, such transformations can create the possibility for discrimination against individuals and groups. But at other times, such transformation can simply violate expectations and create data flows and knowledge that individuals may find inappropriate.

A range of issues surround the transformation of data from sensors or online behaviour into sensitive data that concerns an individual's health, wellbeing or mental state. For example, researchers are demonstrating the potential to turn data from smartphone sensors into data that can be used to predict clinical depression.[153] Other work, such as that in the field of affective computing, attempts to map and predict emotions from different types of sensors, such as blood volume pulse or skin conductivity devices worn on the wrist (e.g. smart watches),[154] photos or videos of faces,[155] gait or posture.[156] Such concerns have led to some

[151] Michael Veale, Max Van Kleek and Reuben Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making' CHI 2018 (ACM Press 2018) https://doi.org/10.1145/3173574.3174014

[152] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, Moritz Hardt (2018) Model Reconstruction from Model Explanations. https://arxiv.org/abs/1807.05185

[153] Farhan and others, 'Behavior vs. Introspection: Refining Prediction of Clinical Depression via Smartphone Sensing Data', *2016 IEEE Wireless Health (WH)* (IEEE 2016).

[154] Rafael Calvo and others, 'Physiological Sensing of Emotion' in Rafael Calvo and others (eds), *The Oxford Handbook of Affective Computing* (Oxford University Press 2015).

[155] H Gunes and M Piccardi, 'Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display' (2009) 39 IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 64.

[156] Rafael Calvo and others, 'Automatic Recognition of Affective Body Expressions' in Rafael Calvo and others (eds), *The Oxford Handbook of Affective Computing* (Oxford University Press 2015); A Kleinsmith, N Bianchi-

researchers proposing new rights entirely, such as the recent report from the Rathenau Instituut, which proposed a 'right not to be measured, analysed or coached' and a 'right to meaningful human contact'.[157] Other researchers have focussed on integrating the existing human rights regime further into concerns around data-driven harms.[158]

Algorithmic systems which measure, count and profile groups of individuals also highlight concerns around what has become known as 'categorical privacy'[159] or 'group privacy'.[160] These issues surround some knowledge that is not (only) private to an individual, but which reveals something about a group. For example, individuals sharing a certain health condition may find that firms gathering data on them are able to detect this health condition through other means, such as through wearable sensors. Communities may share practices such as meeting in certain locations which might allow them to be disclosed and analysed through location data. Genetic data is also commonly described in this way. According to one recent study, 60% of searches for individuals of European-descent based on a set of currently held genetic data from consumer genetic testing companies would result in a third cousin or closer match, which can then allow identification of individuals more specifically using demographic identifiers.[161] The researchers noted that 'a genetic database needs to cover only 2% of the target population to provide a 3rd cousin match to nearly any person'.[162] Individuals' own data can be mined for data about others who did not provide or refused to provide similar data, and the ability to do this is heightened with machine learning technologies.

Finally, there are a range of concerns that algorithmic systems themselves might leak personal data used to train them.[163] A range of attacks have been demonstrated against such systems which 'invert' models: recovering the training data or other private information about the individuals used to train them from the datasets themselves.[164] Such data might also include specific text in the example of language-based algorithmic systems.[165] These properties are of concern to privacy as models become embedded into more and more devices, and traded more freely, and indeed, create challenges in terms of the scope of personal data in data protection law.[166]

The strong privacy provisions in Europe stand in contrast to other regulatory jurisdictions. Given the recent enactment of the GDPR it is not yet fully understood the impact it will have on meaningful explanations and the empowerment of data subjects. Over time, the

---

Berthouze and A Steed, 'Automatic Recognition of Non-Acted Affective Postures' (2011) 41 IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 1027.

[157] Rinie van Est and Joost Gerritsen, *Human rights in the robot age* (Rathenau Instituut 2017).

[158] See eg De Hert P. (2012) A Human Rights Perspective on Privacy and Data Protection Impact Assessments. In: Wright D., De Hert P. (eds) Privacy Impact Assessment. Law, Governance and Technology Series, vol 6. Springer, Dordrecht.

[159] Anton Vedder, 'KDD: The Challenge to Individualism' (1999) 1 Ethics and Information Technology 275.

[160] Linnet Taylor, Luciano Floridi and Bart van der Sloot (eds) *Group Privacy* (Springer 2017).

[161] Yaniv Erlich and others, 'Identity Inference of Genomic Data Using Long-Range Familial Searches' (2018) Science. doi: 10.1126/science.aau4832.

[162] Ibid.

[163] Michael Veale, Reuben Binns, Lilian Edwards, Algorithms that remember: model inversion attacks and data protection law. (2018), http://rsta.royalsocietypublishing.org/content/376/2133/20180083

[164] Matt Fredrikson, Somesh Jha and Thomas Ristenpart, 'Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures' (ACM Press 2015); Reza Shokri and others, 'Membership Inference Attacks Against Machine Learning Models' (IEEE 2017).

[165] Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson and Dawn Song, 'The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets' (2018). https://arxiv.org/abs/1802.08232

[166] Michael Veale, Reuben Binns, Lilian Edwards, Algorithms that remember: model inversion attacks and data protection law. (2018), http://rsta.royalsocietypublishing.org/content/376/2133/20180083

methodological challenges and solutions that the new regulation will provide in the context of privacy will emerge. Still, in the context of privacy and algorithmic decision-making systems, are new rights, such as those proposed by Rathenau Instituut, necessary to protect the privacy of individuals in light of affective computing where data can be extracted and inferred in novel ways? Or are existing regulations and human rights sufficient? What, if any, safeguards are necessary to ensure 'group privacy' is maintained given that information extracted from individuals can reveal private information about a communities/groups?

3.6 Liability

Questions of liability frequently arise in discussions about robotics, self-driving cars and other computational systems which have direct physical effects on the world. Compared to other areas, laws surrounding liability in a range of sectors and applications with difficult challenges have developed internationally over many years. Many contemporary issues concerning algorithmic systems can be seen through the lenses of these rules, and many analogues to algorithmic systems have been proposed: some of which apply given existing rules, and some of which might require statutory change. For the types of harms the liability regimes in Europe already cover, there are few, if any 'gaps' in the law. Instead, core issues surround whether algorithmic harms of salience and concern are effectively covered, and normative questions around whether the distribution of responsibility is a fair one, given the nature of the technologies being discussed.

Some algorithmic technologies are linked clearly to physical consequences, often indirectly. Self-driving vehicles and devices used for diagnostic and treatment in medical settings are the clearest examples of such technologies, and these have the potential to cause physical injury as well as damage to property. Strict liability regimes already apply to particular sectors.[167] Most notably, motorised vehicles, which are not the subject of this report, are generally recognised as subject to near-strict liability, holding the driver to a high standard of care, and holding the driver and/or the manufacturer liable for accidents the vehicle causes. Strict liability would be possible to introduce for deployed algorithmic systems in particular contexts. This could occur for specific applications, such as domestic robots or smart environments in public spaces. However, a general strict liability for algorithmic systems in all contexts, given the broad use of algorithmic systems and the ways in which they blur into many areas of statistics, reasoning and evidence and well as more advanced operational components, risks being excessive, and deploying strict liability into many more areas than are currently covered by it today.

Some liability regimes for do already exist for objects, however. Most jurisdictions recognise liability for some physical objects, particularly in the context of their commercial use. Where algorithmic systems are linked to moveable objects, a range of analogues exist.[168] Some argue that liability for algorithmic systems should be considered similarly to how liability for animals is recognised: a move which would usually require a change in the law. In some countries, this is a negligence-based regime, where an owner can absolve themselves of liability if they take

[167] For example, in English law, this includes keeping dangerous animals, owning an aircraft, or accumulating something unnatural on land which might damage neighbouring land, such as water in a reservoir.
[168] Eric Tjong Tjin Tai, 'Liability for (Semi)Autonomous Systems: Robots and Algorithms' in Vanessa Mak, Eric Tjong Tjin Tai and Anna Berlee (eds), *Research Handbook on Data Science and Law* (Edward Elgar 2018).

sufficient care (Germany); in other countries it is a strict liability regime, where the owner is always liable when damage is caused (France); and some jurisdictions have a mixed approach (England and Wales).[169] According to this view, damage caused by (for example) an autonomous lawnmower would be equivalent to that caused by a sheep, with the owner being liable despite not being personally at fault. Such a view based on current and long-standing legal frameworks stands in contrast to arguments that autonomous and semi-autonomous systems should be granted some aspects of legal personality, which would presumably require a funding regime to allow damages to be recovered from them.

Furthermore, in countries such as France, Belgium and the Netherlands there exists a form of general liability for *moveable objects* besides animals, which requires establishing when an object is causally linked to physical damage. This would be likely to extend to robots insofar as they are considered under the control or mastery of an owner, but some authors disagree on this based on their understanding of the autonomous nature of robots.[170] Many jurisdictions are wary of introducing a liability regime of this sort due to its potentially extensive nature, and similar concerns are likely to occur were a general liability for systems incorporating algorithmic elements to be proposed.

If liability is neither attributed to the system itself (as per legal personality arguments), nor to the customer who purchased it, the obvious candidate for liability is the manufacturer or provider. Indeed, there are many cases in which ordinary product liability law would apply to products which contain elements of algorithmic decision-making, such as digital health devices.

In Europe, product liability laws are harmonised under the Product Liability Directive.[171] These do not apply to algorithmic systems *per se*, but only insofar as they are incorporated into a specific physical product.[172] Producers can be held liable for defects that they could not have known given the state-of-the-art in the area. Product liability is limited to injuries or damage to property, not to economic or other forms of loss, and so only insofar as product with algorithmic components cause those types of damage can they be held liable.

Consumer organisations have advocated extending the scope of European product liability regulations to include digital services.[173] In general, the question of software liability has vexed regulators for decades, owing to the many different ways in which software can be created and distributed, the variety of contexts of application, and the way it is re-purposed and differentially maintained by users.[174] With the advent of Internet-of-Things devices, which generally would fall within the scope of product liability, questions of liability become

[169] Ibid.
[170] Ibid; cf Ryan Calo, 'Robotics and the Lessons of Cyberlaw' (2015) 103 Cal. L. Rev. 513 (on the concept of 'exclusive control').
[171] The Machinery Directive contains provisions specific to some industrial settings. It will not be discussed here.
[172] It is worth noting that product liability has been extended in certain areas, such as to electricity or real estate, however this would require a court to determine that the distribution of algorithmic systems is sufficiently analogous to that of tangible personal property. See Tjong Tjin Tai, 'Liability for (Semi)Autonomous Systems: Robots and Algorithms' in Vanessa Mak, Eric Tjong Tjin Tai and Anna Berlee (eds), *Research Handbook on Data Science and Law* (Edward Elgar 2018).
[173] BEUC-X-2017-039 - 25/04/2017. Review of Product Liability Rules. BEUC Position Paper
[174] Beard, T. Randolph, et al. "Tort liability for software developers: A law & economics perspective." J. Marshall J. Computer & Info. L. 27 (2009): 199.

particularly urgent due to the ease with which such devices can be hacked and used to harm not only the user but also third parties.[175]

In cases where liability regimes such as product liability do apply to providers of algorithmic decision-making, more or less complex forms of liability may apply. Where a single entity has full oversight of the design and deployment of an algorithmic system, such as embedding a pre-built image classification model which does not change, and which the customer has no ability to edit or update, liability may clearly fall with a manufacturer, insofar as such a risk could have been detected in accordance with the state-of-the-art. Liability might take the form of *express* guarantees (where the provider will only be held liable if the system fails to operate according to an expressed standard), or *implied* guarantees (where defects are determined relative to an industry standard).

There may, however, be times where it is not clear which entities are responsible for possible failures. This could occur in situations where data collection, model-building, and deployment are undertaken by one or more different parties, and combine to result in errors which could not be foreseen from any one party's perspective. It might occur when manufacturers argue there are responsibilities for consumers to update the algorithms or software to cope with issues not foreseen at the time of manufacture or with defects. It might also occur if the user uses a system in a way which changes the functionality of the algorithm through learning, and through interaction with an environment, and which it is unclear whether them manufacturer should have anticipated such use or deployment, or at the least should have warned against it.

Product liability is challenged by the distribution of algorithmic systems as cloud services, which, insofar as it appears users receive it as a service rather than as part of a product, would fall out of product liability regimes.[176] This might be the case if, as machine learning systems are generally intended to, there are frequent updates, or if consumers pay for such a machine learning service for their physical device, perhaps on an algorithm marketplace. Insofar as this is the case, and machine learning is marketed as a service, then it is unlikely that such a regime would apply. Machine learning as a service, via traded models or APIs, is an increasingly common product offering.[177]

One approach to dealing with these difficult responsibility issues is to apply insurance mechanisms to them, potentially on a mandatory basis, as is done with vehicle insurance, professional liability insurance, in the case of professions such as lawyers or doctors, and most recently, in countries such as Italy, with drones.[178] Such approaches, were they to be adopted, would likely need simultaneous thought and investment in proper risk assessment mechanisms for the complex, environmentally and societally mediated risks that different robotic or

[175] Butler, Alan. "Products Liability and the Internet of (Insecure) Things: Should Manufacturers Be Liable for Damage Caused by Hacked Devices." U. Mich. JL Reform 50 (2016): 913.
[176] Chris Reed, Elizabeth Kennedy and Sara Nogueira Silva, 'Responsibility, Autonomy and Accountability: Legal Liability for Machine Learning' (2016) Queen Mary University of London, School of Law Legal Studies Research Paper No. 243/2016.
[177] Michael Veale, Reuben Binns, and Lilian Edwards, 'Algorithms that Remember: Model Inversion Attacks and Data Protection Law' (2018) 376 Philosophical Transactions of the Royal Society A 20180083 http://doi.org/10.1098/rsta.2018.0083
[178] Andrea Bertolini and others, 'On Robots and Insurance' (2016) 8 International Journal of Social Robotics 381.

algorithmic systems might cause, else they may result in a dysfunctional market for such insurance products.[179]

Algorithmic systems also have the ability to cause intangible damage, such as that derived from an inaccurate fraud detection or credit rating which might damage an individual's financial reputation, potentially in a defamatory way.[180] These issues have been highlighted in relation to online actors, intermediaries and platforms deploying algorithmic systems at scale. Indeed, such systems deployed by search engines in particular have been central to many accusations and cases around defamation in recent years. These have typically been seen through three different functionalities that search engines provide: providing access to a defamatory site; displaying defamatory snippets in response to searches; and by prompting defamatory searches through an 'autocomplete' feature.[181]

In relation to algorithmic systems indexing sites that have the potential to incur liability due to their defamatory nature, search engines have not escaped liability and gatekeeping duties to the same extent as internet service providers (ISPs) have.[182] In some countries, such as under English defamation law, search engines have been considered as 'secondary publishers' in relation to algorithmically generated search results, meaning that they can avoid liability if they remove material after being given notice of it (under the *innocent dissemination* defence). Search engine liability for defamatory autocomplete searches has also been seen in different ways by different courts. In these cases, platforms cannot always rely on their usual defence, present in many legal regimes, of innocent publication or dissemination, as the algorithm they control is also 'creating' the content (e.g. the search prediction). While in some cases, such as in Germany, search engines have so far been absolved of liability for autocompletion, in other cases, such as in France, Italy and Japan, such cases found against Google.[183]

Several outstanding questions persist: if necessary, what would the optimal liability regime that governs algorithmic decision-making look like? Is a digital services liability framework necessary to cover errors brought on by algorithmic decision-making systems given that this is not covered by the Product Liability Directive? Upon error or failure of an algorithmic decision-making system, who is liable when data collection, model-building, and deployment are undertaken by one or more different parties? If so, would strict liability or negligence-based liability be appropriate, and would this also cover economic and other intangible harms as well as damages to persons and property? What are the merits and drawbacks of establishing legal personalities to cover algorithmic decision systems? Finally, would an insurance market be an appropriate mechanism to address the any liability gap that exists for digital services, and what other mechanisms are available?

[179] Ibid.
[180] Chris Reed, Elizabeth Kennedy and Sara Nogueira Silva, 'Responsibility, Autonomy and Accountability: Legal Liability for Machine Learning' (2016) Queen Mary University of London, School of Law Legal Studies Research Paper No. 243/2016.
[181] Uta Kohl, 'Google: The Rise and Rise of Online Intermediaries in the Governance of the Internet and beyond (Part 2)' (2013) 21 International Journal of Law and Information Technology 187.
[182] Uta Kohl, 'The Rise and Rise of Online Intermediaries in the Governance of the Internet and beyond – Connectivity Intermediaries' (2012) 26 International Review of Law, Computers & Technology 185.
[183] Ibid.

3.7 Intermediate findings

The academic debate section attempts to synthesise the latest research from the academic literature, and present the emerging questions related to algorithmic decision-making that remain to be answered through the perspectives of fairness and equity, transparency and scrutiny, accountability, robustness and resilience, privacy, and liability. In its current form it relies heavily on desk research and therefore more concrete conclusions will be drawn after the initial consultation period.

Despite this, the concerns cited throughout the academic debate around algorithmic systems touch upon a huge array of areas of societal concern. Some of these are extensions of old challenges with added complexity from the changing and distributed nature of these technologies: such as liability concerns, or societal discrimination. Others, however, seem newer, such as the transformation of mundane data into private or sensitive data, or the new and unusual ways in which technologies might fail or be compromised. Scholars from a wide variety of disciplines have weighed in on how these issues play out in a technical sense, and how they see these issues in relation to governance, existing social and policy problems, societal framing and involvement in technological innovation, legal and regulatory frameworks and ethics. In many cases, these issues are not new, but they are reaching a salience and importance they did not previously have.

There are several outstanding questions in the research literature relating to algorithmic systems which have had little focus upon, or which remain open. For example, in the context of fairness, do citizens and businesses feel that systems which have been 'debiased' are more legitimate on the ground, and do such systems actually mitigate or reduce inequalities in practice? Related to robustness and resilience, to what extent can different algorithmic systems be fooled in practice and within business models that exist today, and how great is the risk to society that they may be able to be manipulated in ways that are difficult to defend against? What methods of transparency, particularly to society rather than just to individuals, might promote effective oversight over the growing number of algorithmic systems in use today?

The academic debate has also highlighted inter-perspective tensions and it remains to be seen, for example, whether machine learning systems can be audited for difficult-to-spot discrimination, thereby addressing transparency and accountability, without collecting highly sensitive data about individuals, which may be detrimental to individual and group privacy?

Some of these, and other questions that have been raised throughout this section will be explored further through sector/application-specific case studies which will form part of the evidence-base from which policy solutions may be designed. Beyond these future questions, it seems unlikely that a single policy solution or approach will deal with all, or even most of those challenges currently identified. In order to address all of them, and to manage the trade-offs that arise, a layered variety of approaches are likely to be required. Civil society and industry have already begun to develop initiatives and design technical tools to address some the issues raised throughout this section. It is to these that the following sections of this report now turn.

4. Initiatives from industry, civil society and other multi-disciplinary organisations

4.1 Overview

This section aims to complement the analytical literature review presented above by **providing an outline of initiatives tackling the key challenges facing algorithmic decision-making**. In particular, this section presents initiatives from civil society, industry and other relevant organisations. Whilst acknowledging that there is not a single way to group or cluster the different initiatives in this context, the initial research undertaken has identified four main types of relevant initiatives, namely:

- **Standardisation efforts:** given the stature of key stakeholders in this domain, three examples are presented, including projects currently being carried out by the IEEE and the ISO;
- **Codes of conduct, ethical principles and ethics frameworks for AI and algorithmic decision-making:** many organisations have engaged in this type of initiative across industry, academia and elsewhere. Here, we present eleven examples from the most prominent and relevant sources in the field;
- **Working groups and committees:** four examples of established initiatives that conduct research and foster collaboration and an open dialogue are presented;
- **Policy and technical tools**: this category of initiatives primarily includes tools developed and promoted by academic projects, as well as in industry and other research organisations to tackle specific challenges. Here, we present six examples.

The following sub-sections present a summarised outline of the different initiatives found in the abovementioned categories, highlighting if they engage with the six concepts of **fairness and equity** (F&E), **transparency and scrutiny** (T&S), **accountability** (Acc.), **robustness and resilience** (R&R), **privacy** (Priv.) and **liability** (Liab.). The below summary table presents an introduction to each initiative and indicates the concepts with which each initiative engages.

The initiatives outlined and discussed below are not limited to algorithmic decision-making, but also include closely related concepts such as 'automated [decision-making] systems', 'autonomous systems', 'AI / intelligent systems / cognitive systems[184]'.

[184] In the instances where initiatives refer to intelligent decision-making systems.

Table 1: Summary of initiatives vs. challenges facing algorithmic decision-making

Initiative	F&E	T&S	Acc.	R&R	Priv.	Liab.
Standardisation efforts						
IEEE AI standardisation efforts, including the *Ethically Aligned Design* (EAD) initiative and the IEEE P7000™ Working Groups	■	■	■		■	
ISO JTC1 / SC42 committee, which comprises a working group on foundational standards and study groups focused on: i) computational approaches and characteristics of AI; ii) trustworthiness; and iii) use cases and applications	■	■		■	■	
China White Paper on standardisation, which highlights specific concerns related to safety and security, ethics and privacy and presents China's thoughts on standardisation	■	■	■	■	■	■
Codes of conduct, ethical principles and ethics frameworks for AI and algorithmic decision-making						
SIIA ethical principles. Focusing on responsible data use, SIIA's principles aim to ensure organisations can demonstrate to policy-makers and the public that they are responsible		■	■			
ITI Council policy principles aim to promote responsible development and use of AI through collaboration, primarily covering the issues of interpretability and liability	■			■		■
Partnership on AI and principles in industry presents some of the many industry efforts to establish an AI ethics framework, including the most prominent example: the 'Partnership on AI'	■	■	■	■	■	
NESTA public sector principles. Highlighting the particularly problematic use of opaque AI the public sector, NESTA has drafted 10 principles for public sector use of algorithmic decision-making	■	■	■	■	■	■
Digital Catapult AI ethics framework. Digital Catapult has developed an ethics framework, which it has invited AI companies to test, in particular the start-ups it supports	■	■	■		■	
Algorithmenethik, an initiative aimed at first understanding what makes a successful professional ethical code and subsequently designing an #algorules catalogue of quality criteria	■	■	■		■	■
Future of Life Institute Asilomar principles are a wide-ranging set of 23 principles focusing on three areas: i) research issues; ii) ethics and values; and iii) longer-term issues	■	■	■	■	■	
Montreal declaration for responsible AI development. Established at the Forum on Socially Responsible Development of AI, this initiative details seven principles for responsible AI development	■	■	■		■	
FAT/ML principles and social impact statement. Through these two initiatives, FAT/ML aims to support developers and product managers develop publicly accountable algorithmic systems	■	■	■	■	■	■
ICDPPC declaration on ethics and data protection, which presents guiding principles and has the core aim of preserving human rights in AI development. The ICDPPC has also established a permanent working group	■	■	■		■	

Initiative	F&E	T&S	Acc.	R&R	Priv.	Liab.
Cowls & Floridi ethical framework. An academic analysis of prominent ethics frameworks for AI development that concludes with its own framework based on the existing examples and established bioethics principles	■	■	■	■	■	
Working groups and committees conducting research and fostering collaboration and an open dialogue						
Research and advocacy in AI. This example presents a wide-range of civil society efforts to collaborate on tackling the challenges facing algorithmic decision-making, including initiatives by: TransAlgo, algodiv, AlgorithmWatch, MIRI and FAT/ML	■	■	■	■	■	
World Wide Web Foundation initiative. Under its main aim of Establishing the open Web as a basic right and a public good, the Foundation has conducted research on the opportunities and risks of AI	■	■	■			
Policy and technical tools						
AI NOW impact assessments and accountability policy toolkit. These initiatives focus on understanding and presenting the social implications of AI across: i) rights and liberties; ii) labour and automation; iii) bias and inclusion; and iv) safety and critical infrastructure	■	■	■			■
UnBias Fairness toolkit, which aims to promote awareness and stimulate public dialogue on algorithmic fairness	■					
CDT Digital decisions tool aims to support engineers and product managers in mitigating the risks of designing unfair, discriminatory and harmful decision-making algorithmic systems	■			■		
Accenture / ATI algorithmic fairness tool. This tool scrutinises the input data for an algorithm and aims to uncover any unfairness or discrimination	■					
DTL technical programme and tools, which aim to connect key stakeholders to co-develop solutions that ensure transparency and support the practical development of useful tools	■	■		■	■	
EthicsToolkit focuses on empowering and supporting the public sector to understand and mitigate the risks of algorithmic decision-making	■	■	■		■	

4.2 Standardisation efforts

IEEE's AI standardisation efforts – Ethically aligned design and beyond

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS Ethics)[185] has recently published an updated version of its report on *Ethically Aligned Design (EAD)*[186]. The mission of this initiative, as outlined in the report is "to ensure every technologist is educated, trained, and empowered to prioritise ethical considerations in the design and development of autonomous intelligent systems"[187]. With the engagement and participation of more than 250 stakeholders[188], the initiative is promoting global discussions and debates including: "advocacy for the positive impact, as well as warnings, based on the potential harm to privacy, discrimination, loss of skills, economic impacts, security of critical infrastructure, and the long-term effects on social well-being".

In promoting debate around the ethical principles for the design of autonomous systems, the IEEE aims to identify and find broad consensus on pressing ethical and social issues and provide recommendations regarding development and implementations of these technologies. Specifically, the initiative has outlined the following general principles as initial guidelines for the ethical design, development, and implementation of intelligent and autonomous systems[189]:

- **Human Rights:** ensure they do not infringe on internationally recognized human rights
- **Well-being:** prioritise metrics of well-being in their design and use
- **Accountability:** ensure that their designers and operators are responsible and accountable
- **Transparency:** ensure they operate in a transparent manner
- **Awareness of misuse:** minimize the risks of their misuse

In addition to the work being carried out through the *Ethically Aligned Design* initiative, the IEEE has also established the IEEE P7000™ Working Groups. These aim to co-develop the standards for the future of ethical intelligent and autonomous technologies on several fronts. They include the following[190]:

- P7000. Model process for addressing ethical concerns during system design
- P7001. Transparency of autonomous systems
- P7002. Data privacy process
- P7003. Algorithmic bias considerations
- P7004. Standard on child and student data governance
- P7005. Standard on employer data governance
- P7006. Standard on personal data AI agent working group
- P7007. Ontological standard for ethically driven robotics and automation systems
- P7008. Standard for ethically driven nudging for robotic, intelligent and autonomous systems
- P7009. Standard for fail-safe design of autonomous and semi-autonomous systems

[185] See https://ethicsinaction.ieee.org/
[186] See https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf
[187] Ibid.
[188] Ibid.
[189] Ibid.
[190] See https://ethicsinaction.ieee.org/

- P7010. Wellbeing metrics standard for ethical artificial intelligence and autonomous systems
- P7011. Standard for the process of identifying and rating the trust-worthiness of news sources
- P7012. Standard for machine readable personal privacy terms
- P7013. Inclusion and application standards for automated facial analysis technology

Since the release of *Ethically Aligned Design* version 1 (EADv1) in December 2016, the IEEE has grown its stakeholder engagement to capture a more balanced global perspective. Notably, the IEEE recognised a 'Western' dominance in EADv1 and, as a result, expanded the Initiative to add members from China, Korea, Japan, Brazil, Mexico, Russia, Iran, Thailand and Israel. The membership of the Initiative grew from 100 to 250. Furthermore, the EADv1 Executive Summary has been translated into nine languages: Arabic, Chinese, Japanese, Korean, Persian (Farsi), Portuguese, Russian and Thai.

Subsequently, in December 2017, the IEEE published version 2 (EADv2), which, to illustrate the engagement of stakeholders, received at least 72 responses (including many notable parties) via a public request for input. Furthermore, alongside EADv2, the IEEE published Regional Reports detailing initiatives in the field of A/IS Ethics from thought leaders in Japan, Hong Kong, Asian Fintech more generally, Russia and Israel.[191]

The International Organisation for Standardisation (ISO) JTC1 / SC42 committee

Through its Joint Technical Committee (JTC 1), the ISO recognises that AI is a field which mobilises a diverse landscape of stakeholders including those involved in research, academia, industry, policy makers, ethics advocates and more. Moreover, the application areas of AI technology are equally as diverse and numerous. It also recognises that the structured deployment and adoption of AI, as a cross-sectoral technology (i.e. or group of technologies), requires the development of standards[192].

To this end, the ISO has established the JTC 1/SC 42 committee, which contains one working group and three study groups with a diverse work programme, namely[193,194]:

- SC42/WG1 – The **foundational standards working group**, which will take on two ongoing standardisation projects: *Artificial Intelligence Concepts and Terminology ISO/IEC AWI 22989* and *Framework for Artificial Intelligence Systems Using Machine Learning ISO/IEC AWI 25053;*
- SC42/SG1 – The **computational approaches and characteristics of artificial intelligence study group**, which will study: (i) different technologies (including machine learning algorithms) used by AI systems including their characteristics and properties; (ii) existing specialised AI systems, such as natural-language processing or computer vision systems, with the objective of understanding their computational architectures and approaches; and (iii) industry practices, processes and methods for the application of AI systems;

[191] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Regional Reports on A/IS Ethics, December 2017
[192] See https://jtc1info.org/jtc1-press-committee-info-about-jtc-1-sc-42/
[193] Ibid.
[194] https://www.linkedin.com/pulse/overview-artificial-intelligence-ethics-regulations-guttmann/

- SC42/SG2 – The **trustworthiness study group**, which will:
  - "*Investigate approaches to establish trust in AI systems through transparency, verifiability, explainability, controllability, etc.;*
  - *Investigate engineering pitfalls and assess typical associated threats and risks to AI systems with their mitigation techniques and methods;*
  - *Investigate approaches to achieve AI systems' robustness, resiliency, reliability, accuracy, safety, security, privacy, etc.;*
  - *Investigate types of sources of bias in AI systems with a goal of minimization, including but not limited to statistical bias in AI systems and AI aided decision-making*";
- SC42/SG3 – The **use cases and applications study group**, which will:
  - *"Identify different AI application domains (e.g., social networks and embedded systems) and the different context of their use (e.g., fintech, health care, smart home, and autonomous cars);*
  - *Collect representative use cases;*
  - *Describe applications and use cases using the terminology and concepts defined in ISO/IEC AWI 22989 and ISO/IEC AWI 23053 and extend the terms as necessary;*
  - *Develop new work item proposals as appropriate and recommend placement".*

Additionally, as horizontal issues across all four groups, the ISO committee is focused on:

- the **societal concerns and ethical considerations related to the use of AI**, noting, for example, that issues such as algorithmic bias, eavesdropping and safety directives in industrial AI fall within its remit; and
- the **role of big data and its interplay with artificial intelligence**, noting that it is difficult to envisage applications where one technology is present without the other.[195]

Created in 2017, the JTC 1/SC 42 committee comprises 22 members and 8 observers from standardisation institutes across the globe. In 2018, the committee published its first two standards:

- ISO/IEC TR 20547-2:2018: Information Technology – Big data reference architecture – **Part 2: Use cases and derived requirements**. The first standard developed by the committee provides examples of big data use cases and derives application domains and technical considerations from those use cases.
- ISO/IEC TR 20547-5:2018: Information Technology – Big data reference architecture – **Part 5: Standards roadmap**. The second standard details the status and roadmap of standards relating to big data.

Furthermore, the JTC 1/SC 42 committee is currently developing a further five standards relating to:

i) Big data overview and vocabulary;
ii) Big data reference architecture – Part 1: Framework and application process
iii) Big data reference architecture – Part 3: Reference architecture
iv) Artificial Intelligence – Concepts and terminology
v) Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)

[195] ISO news item (R. Bartram), The new frontier for artificial intelligence, 18 October 2018. Last accessed on 06.11.2018 at: https://www.iso.org/news/ref2336.html

China's 'White Paper on Artificial Intelligence Standardisation' **(2018)**

Developed by more than 30 academic and industry organisations under the guidance of the Chinese Electronics Standards Institute, the White Paper on AI Standardisation[196] presents China's thoughts on standardization in AI. In particular, this White Paper recognizes the importance of standardization of AI with a dedicated sub-section.

Furthermore, the White Paper detail China's specific concerns in relation to:

- **Safety and security:** the White Paper discusses the notions of algorithmic harm and the pervasiveness of the ability to develop AI algorithms and thus the difficulty of controlling algorithmic development;
- **Ethics:** the White Paper introduces three key ethical principles on which consensus has been reached: (i) the principles of human interests (i.e. AI should benefit human welfare); (ii) the principle of liability (i.e. a clear system of liability and compensation needs to be established, building on principles of transparency, equal rights and responsibility); and (iii) the "consistency of rights and responsibilities" principle (i.e. the legal framework should clearly define the balance between oversight and review of the data and algorithms in use and the protection of intellectual property rights for developers); and
- **Privacy issues:** the White Paper notes that, given that recent AI developments are based on large amounts of data, a clear definition of privacy in this context is required.

4.3 Codes of conduct, ethical principles and ethics frameworks for AI and algorithmic decision-making

Software and Information industry Association (SIIA) – *Ethical Principles for AI and Data Analytics*

Through its publication 'Ethical Principles for Artificial Intelligence and Data Analytics'[197], the SIIA highlights that organisations should not be indifferent to how and by whom the models they develop are used, as well as how the benefits of their new analytical services are distributed.[198] Specifically, the positioning of the SIIA in the ethics of the algorithmic decision-making arena is that "organisations can meet their ethical obligations and persuade policymakers and the public that they are responsible users of data analytics only if they have policies and procedures in place for ethical review, publicly available ethical principles that they adhere to and a transparent communication programme that allows them to describe their policies, procedures and principles in an accountable way".

In this context, the SIIA proposes a framework of responsible data use, including ethical principles for institutions to use and rely on to assess different models and data. This work is analytically summarised by SIIA through different discussions, including: traditional ethical

[196] See https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-excerpts-chinas-white-paper-artificial-intelligence-standardization/

[197] See http://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990

[198] See https://www.linkedin.com/pulse/overview-artificial-intelligence-ethics-regulations-guttmann/

frameworks; general principles for data practices; principles of human experimentation; core values for a shared ethical framework; and the United Nations principles on business and human rights. In addition to providing this framework and principles, the framework seeks to "promote further discussion among policymakers, organisations developing and using data, activists, scholars, ethicists, and civil society"[199].

It is also worth noting SIIA's explicit terminological statements regarding "data analytics system", "advanced analytical model" and "data practices"[200].

Information technology Industry Council (ITI)

The Information Technology Industry Council (ITI) represents the technology sector's leading companies and has recently been promoting increased collaboration among stakeholders across public and private sectors. The ITI acknowledges the need to develop multi-disciplined dialogues with governments and other interested parties in order to ensure that AI is able to deliver its greatest positive potential[201]. In fact, in its publication on 'AI Policy Principles', the ITI explicitly refers to algorithmic decision-making in its principles on *Promoting Responsible Development and Use*[202], as outlined below:

- ***Interpretability:*** *We are committed to partnering with others across government, private industry, academia, and civil society to find ways to mitigate bias, inequity, and other potential harms in automated decision-making systems. Our approach to finding such solutions should be tailored to the unique risks presented by the specific context in which a particular system operates. In many contexts, we believe tools to enable greater interpretability will play an important role.*
- ***Liability of AI Systems Due to Autonomy:*** *The use of AI to make autonomous consequential decisions about people, informed by – but often replacing decisions made by – human-driven bureaucratic processes, has led to concerns about liability. Acknowledging existing legal and regulatory frameworks, we are committed to partnering with relevant stakeholders to inform a reasonable accountability framework for all entities in the context of autonomous systems.*

Furthermore, although not specifically referencing algorithmic decision-making, the ITI's AI Policy Principles discuss the need to ensure: safety and controllability in the design of AI systems; and responsible use of data to ensure its integrity and test for potential bias.

---

[199] See http://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990

[200] Specifically, "data analytics system" is defined by the SIIS as any method of processing or analyzing data, traditional or advanced, that reveals insights relevant for decision-making. "Advanced analytical model" refers to any of the newer techniques of data analysis including artificial intelligence and machine learning that rely on large quantities of data to train algorithms to produce a desired output. "Data practices" refers collectively to the process of data collection, analysis and use of information for decision-making purposes. See http://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990

[201] See https://www.itic.org/public-policy/ITIAIPolicyPrinciplesFINAL.pdf

[202] Ibid.

AI Ethics Principles in the technology industry and the Partnership on AI

Whilst not providing a specific definition for AI systems, nor explicitly associating them to decision-making systems, IBM characterises the purpose "of AI and cognitive systems developed and applied by IBM is to augment human intelligence and [...] enhance and extend human capability, expertise and potential". This positioning is further outlined in detail throughout its *Principles for Trust and Transparency*[203]. In addition, Adam Cutler, Milena Pribić and Lawrence Humphrey, from IBM Design, have recently published a multi-disciplinary practical guide for AI designers and developers entitled *Everyday Ethics for Artificial Intelligence*. A recent and ongoing initiative, this guide was organised around 5 key areas / principles, namely[204],[205]:

- ***Accountability:*** *AI designers and developers are responsible for considering AI design, development, decision processes, and outcomes.*
- ***Value Alignment:*** *AI should be designed with consideration of the norms and values of your user group.*
- ***Explainability:*** *AI should be designed for humans to easily perceive, detect, and understand its decision process.*
- ***User Data Rights:*** *AI should be designed to protect user data and preserve the user's power over access and uses.*
- ***Fairness:*** *AI should be designed to minimize bias and promote inclusive representation.*

Furthermore, IBM Research has partnered with the MIT Media Lab to develop an AI recommendation technique which is capable of optimising results according to user preferences while staying conformant to behavioural and ethics constraints[206].

Generally, large technology and online platform companies, such as Google, Facebook, Microsoft and SAP, have reportedly created dedicated teams to embed ethical design and development of AI systems in their organisations[207,208,209]. This has coincided with these companies publishing AI ethics principles. Google, for example, has recently published its principles for the design and development of AI systems, which are as follows[210]:

- Be socially beneficial;
- Avoid creating or reinforcing unfair bias;
- Be built and tested for safety;
- Be accountable to people;
- Incorporate privacy design principles;
- Uphold high standards of scientific excellence;
- Be made available for uses that accord with these principles.

---

[203] See https://www.ibm.com/blogs/policy/trust-principles/
[204] See https://medium.com/design-ibm/everyday-ethics-for-artificial-intelligence-75e173a9d8e8
[205] The full guide can be accessed at https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf
[206] See https://venturebeat.com/2018/07/16/ibm-researchers-train-ai-to-follow-code-of-ethics/
[207] See https://www.wired.com/story/tech-firms-move-to-put-ethical-guard-rails-around-ai/
[208] See https://www.accenture.com/gb-en/company-responsible-ai-robotics
[209] See https://www.sap.com/products/leonardo/machine-learning/ai-ethics.html
[210] A more detailed account of these principles can be accessed at https://www.blog.google/technology/ai/ai-principles/

Despite being generally well-received, a few concerns with these principles were noted by the international community[211]. One of the main concerns is that Google, as well as other technology companies who have published similar principles for the ethical design of AI systems, has not explicitly committed to the type of independent, informed and transparent review which would be ideal for ensuring the principles are always applied and applied well[212].

In addition to the publication of codes of conduct and principles for the ethical design and development of AI and algorithmic decision-making systems, the idea of technology companies adhering to a 'digital Hippocratic oath' has recently been proposed[213]. In a book, recently released by Microsoft[214], it is proposed that "it could make sense" to bind coders to a pledge like that taken by physicians to "first do no harm". Microsoft has also reported that it is working on ensuring that the AI systems it develops, and their underlying training and input data, are "accurate and unbiased"[215].

Generally, technology companies have also joined international and multi-disciplinary groups to debate and find solutions to the impacts of AI and algorithmic decision-making in society. The **Partnership on AI** is a prime example of this[216]. In fact, the Partnership on AI aims to shape best practices, research, and public dialogue about the benefits of AI for people and society. Through engagement with stakeholders from industry, civil society and academia it also aims to co-create solutions in this field according to 6 thematic pillars[217]:

- Safety-critical AI;
- Fair, transparent, and accountable AI;
- AI, labour and the economy;
- Collaboration between people and AI systems;
- Social and societal influences of AI;
- AI and social good.

Considering these thematic pillars, the Partnership on AI aims to undertake the following steps to support the ethical development of AI:

i) **engage experts** across many disciplines including psychology, philosophy, economics, finance, sociology, public policy and law.
ii) **engage stakeholders**, meaning users, developers and all industry sectors potentially impacted by AI, noting in particular healthcare, financial services, transportation, commerce, manufacturing, telecommunications and media.
iii) **support objective third-party studies**, in particular on best practices for the safety, ethics, fairness, inclusiveness, trust and robustness of AI research, applications and services.
iv) **develop learning materials** on current and future AI trends.

Initiated in late 2016 by six companies (Apple, Amazon, Google and DeepMind, Facebook, IBM and Microsoft), the Partnership on AI now combines the expertise and interest of more than

[211] See https://www.eff.org/deeplinks/2018/06/how-good-are-googles-new-ai-ethics-principles
[212] Ibid.
[213] See https://www.wired.com/story/should-data-scientists-adhere-to-a-hippocratic-oath/
[214] See https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/
[215] See https://www.microsoft.com/en-us/research/blog/ai-getting-smarter-microsoft-researchers-ensure-ai-accuracy/
[216] A full list of partners can be accessed at https://www.partnershiponai.org/partners/
[217] See https://www.partnershiponai.org/about/#our-work

70 stakeholders, of which more than 50% represent non-profit organisations, spanning 10 countries covering Europe, North America, Asia and Oceania.

NESTA – Principles for public sector use of algorithmic decision-making

NESTA has claimed that the case for creating more robust codes of conduct regarding the use of algorithmic decision-making is stronger for the public sectors. It uses the rationale provided by Robert Brauneis and Ellen P. Goodman[218] to justify this assessment: "In the public sector, the opacity of algorithmic decision-making is particularly problematic both because governmental decisions may be especially weighty, and because democratically-elected governments bear special duties of accountability".

In this context, NESTA provided, in February 2018, a working draft of 10 principles for public sector use of algorithmic decision-making, as outlined below[219]:

- *Every algorithm used by a public sector organisation should be accompanied with a **description of its function, objectives and intended impact**, made available to those who use it;*
- *Public sector organisations should publish details **describing the data on which an algorithm was (or is continuously) trained**, and the **assumptions used** in its creation, together with a risk assessment for mitigating potential biases;*
- *Algorithms should be **categorised on an Algorithmic Risk Scale** of 1-5, with 5 referring to those whose impact on an individual could be very high, and 1 being very minor;*
- *A **list of all the inputs used** by an algorithm to make a decision should be published.*
- ***Citizens must be informed** when their treatment has been informed wholly or in part by an algorithm;*
- *When using third parties to create or run algorithms on their behalf, public sector organisations should **only procure from organisations able to meet Principles 1-6**;*
- *A named member of **senior staff (or their job role) should be held formally responsible** for any actions taken as a result of an algorithmic decision;*
- *Public sector organisations wishing to adopt algorithmic decision-making in high risk areas should sign up to a **dedicated insurance scheme** that provides compensation to individuals negatively impacted by a mistaken decision made by an algorithm;*
- *Public sector organisations should commit **to evaluating the impact of the algorithms** they use in decision-making and publishing the results.*

Digital Catapult – AI Ethics Framework

Digital Catapult's Machine Learning Garage[220] has released its first ethics framework and invited AI companies to test it as a means of integrating ethical practice into the development of AI and machine learning technologies[221,222]. In contrast to NESTA's principles, this framework

[218] See https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3012499
[219] For a more detailed explanation of each principle, see https://www.nesta.org.uk/blog/10-principles-for-public-sector-use-of-algorithmic-decision-making/
[220] See https://www.migarage.ai/
[221] See https://www.digicatapult.org.uk/news-and-views/press/digital-catapult-adoption-of-ethics-in-ai/
[222] See https://www.migarage.ai/ethics-framework/

is targeted at start-ups who are a part of the Machine Intelligence Garage to have access to computation power and expertise to accelerate technological development. Specifically, the aim of the framework is to "help companies to characterise the ethical opportunities and potential risks associated with their business and / or technology and be open and clear both internally and externally about how these are evaluated and managed"[223].

The framework, devised by a dedicated Ethics Committee[224], comprises the following seven practical concepts, which aim to guide the ethical use of AI by the participating start-ups:

- Be clear about the benefits of your product or service.
- Know and manage your risks.
- Use data responsibly.
- Be worthy of trust.
- Promote diversity, equality and inclusion.
- Be open and understandable in communications.
- Consider your business model.

As of November 2018, the Machine Learning Garage supports a cohort of 23 start-ups.

Algorithmenethik

In its study "professional ethics code for the design of algorithms", Algorithmenethik takes a different approach within this set of initiatives. Instead of detailing an ethics code, it aims to establish success factors for the development of a professional ethics code and how this translates into action points for the design of algorithmic ethics frameworks. The following details the ten success factors and their meaning for algorithmic design:

- **Historic tradition**: Many professional ethics codes root in long-standing occupational tradition and evolution over several generations. For the field of algorithmic design this means that normative traditions and ethical principles of algorithms need to be established.
- **Personal cause:** Members of the occupational group need to be motivated and have a socio-ethical understanding. For the field of algorithmic design this means that existing initiatives on the responsible design of algorithms need to support the development of the principles.
- **Grounding:** Occupational ethics need to be understood both within and outside of the occupational group and should be the subject of discussion and exchange. The group surmises that, for the field of algorithmic design, this means that digital education is necessary to ensure understanding outside the industry.
- **Homogeneity of the occupational group:** The occupational group needs to be homogeneous. For the field of algorithmic design this means that the ethics code needs to be formulated as specifically as possible.
- **Raising awareness through education:** The ethical principles need to be part of the educational system. For the field of algorithmic design this means that established professional ethics codes of related occupational fields need to be elaborated upon.

[223] Ibid.
[224] See https://www.migarage.ai/ethics-committee/

- **Institutionalisation of the ethics code through professional organisations:** For the field of algorithmic design this means that the occupational field in relation to algorithmic design needs to be established.
- **Sanctions:** For the field of algorithmic design this means that certain behaviours need to be blacklisted and punished through fines and disbarment.
- **Material background:** Use of financial resources for the development and updating of the ethics code. For the field of algorithmic design this means that sufficient resources need to be available to ensure participation of all persons concerned, as well as the participation of experts.
- **Academic reflection:** For the field of algorithmic design this means that specific challenging situations shall be subject to academic assessment. An understanding of theory and practice is essential.
- **Long-term engagement:** This includes the long-term efforts for continuous development of the ethics code. For the field of algorithmic design this means that sustainable and multi-disciplinary efforts are necessary.

To build on this, Algorithmenethik, which is funded by the Bertelsmann Foundation, in combination with think tank iRights.Lab, has been developing a catalogue of quality criteria for algorithmic processes – the **#algorules process**. At present, the group is conducting an extensive consultation exercise with experts from science, business, politics, civil society and the media, among others.

Future of Life Institute – Asilomar AI Principles

In conjunction with the 2017 Asilomar conference, undertaken in January 2017, the Asilomar AI Principles were developed. In total, there are 23 principles, separated into three areas, as follows:

- **Research issues:**
  - The goal of AI research should be to create not undirected intelligence, but beneficial intelligence;
  - Investments in AI should be accompanied by funding for research on ensuring its beneficial use;
  - There should be constructive and healthy exchange between AI researchers and policy-makers.
  - A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
  - Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.
- **Ethics and Values:**
  - AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
  - If an AI system causes harm, it should be possible to ascertain why.
  - Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

  - Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
  - Highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation.
  - AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
  - People should have the right to access, manage and control the data they generate, given AI systems' power to analyse and utilize that data.
  - The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
  - AI technologies should benefit and empower as many people as possible.
  - The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
  - Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
  - The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
  - An arms race in lethal autonomous weapons should be avoided.
- **Longer-term issues:**
  - There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
  - Advanced AI could represent a profound change in the history of life on Earth and should be planned for and managed with commensurate care and resources.
  - Risks posed by AI systems must be subject to planning and mitigation efforts commensurate with their expected impact.
  - AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
  - Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

In contrast to most other work on developing an ethics framework for AI development, the Future of Life Institute collects data on the commitments made to the Asilomar principles, reporting that 1273 AI/Robotics researchers and 2541 others have become signatories of the principles.

Montreal Declaration for a Responsible Development of Artificial Intelligence

Announced in November 2017 at the conclusion of the Forum on the Socially Responsible Development of AI, the Montreal Declaration details the following seven principles:

- **Well-being:** The development of AI should ultimately promote the well-being of all sentient creatures.
- **Autonomy:** The development of AI should promote the autonomy of all human beings and control, in a responsible way, the autonomy of computer systems.

- **Justice:** The development of AI should promote justice and seek to eliminate all types of discrimination, notably those linked to gender, age, mental / physical abilities, sexual orientation, ethnic / social origins and religious beliefs.
- **Privacy:** The development of AI should offer guarantees respecting personal privacy and allowing people who use it to access their personal data as well as the kinds of information that any algorithm might use.
- **Knowledge:** The development of AI should promote critical thinking and protect us from propaganda and manipulation.
- **Democracy:** The development of AI should promote informed participation in public life, cooperation and democratic debate.
- **Responsibility:** The various players in the development of AI should assume their responsibility by working against the risks arising from their technological innovations.

Fairness, Accountability and Transparency in Machine Learning (FAT/ML): Principles and Social Impact Statement

FAT/ML aims to bring together researchers and practitioners concerned with fairness, accountability and transparency in ML. Key outputs of this collaborative group have been:

- **Principles** for Accountable Algorithms; and
- **Social Impact Statement** for Algorithms.

Through the former output, FAT/ML aimed to support developers and product managers to 'design and implement algorithmic systems in publicly accountable ways'[225]. FAT/ML elaborate accountability, in this instance, as an 'obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms.'[226] In addition, FAT/ML specifies the following five guiding principles:

- **Responsibility** – ensure routes of redress are available and visible.
- **Explainability** – ensure algorithmic decisions and data can be explained to stakeholders in non-technical terms.
- **Accuracy** – work to understand and improve the accuracy of the algorithm and its data sources.
- **Auditability** – permit third parties to review the algorithm.
- **Fairness** – ensure discrimination or unjust impacts do not exist when comparing across different demographics.

Interestingly, FAT/ML notes that two important principles – privacy and the impact of human experimentation – have been purposefully excluded, as they are covered elsewhere, linking to the OECD's privacy principles[227] and the US Department for Health and Human Services 'Belmont Report on Ethical Principles and Guidelines for the Protection of Human Subjects of Research'[228].

[225] See http://www.fatml.org/resources/principles-for-accountable-algorithms
[226] Ibid.
[227] See http://oecdprivacy.org/
[228] See https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/

In addition to the development of the principles themselves, FAT/ML also proposes that algorithm developers create a **Social Impact Statement**, to be made public and detail how the developer has considered the above principles.

Data Protection and Privacy Commissioners – Declaration on ethics and data protection in artificial intelligence

At the 40th International Conference of Data Protection and Privacy Commissioners (ICDPPC), conducted in October 2018, the Conference endorsed the following guiding principles, with the core aim of preserving human rights in AI development:[229]

- Artificial intelligence and machine learning technologies should be designed, developed and used in respect of fundamental human rights and in accordance with the **fairness principle**.
- **Continued attention and vigilance**, as well as accountability, for the potential effects and consequences of, artificial intelligence systems should be ensured.
- AI **systems transparency and intelligibility** should be improved, with the objective of effective implementation.
- As part of an overall "ethics by design" approach, artificial intelligence systems should be **designed and developed responsibly**, by applying the principles of **privacy by default and privacy by design**.
- **Empowerment of every individual** should be promoted, and the exercise of individuals' rights should be encouraged, as well as the creation of opportunities for public engagement.
- **Unlawful biases or discriminations** that may result from the use of data in artificial intelligence should be reduced and mitigated.

Alongside these guiding principles, the ICDPPC called for the development of common governance principles on AI and established a permanent working group on AI development – the working group on Ethics and Data Protection in AI.[230]

The ICDPPC is currently requesting feedback on these principles via an open public consultation.[231]

Ethical Framework for a Good AI Society **(Cowls and Floridi)**

Considering all of the above examples, as well as the IEEE principles (detailed under the Standardisation efforts heading) and others[232], it is clear that a large number of outputs have been produced by notable individuals and organisations on the theme 'Codes of conduct, ethical principles and ethics frameworks for AI development'.

---

[229] ICDPPC, Declaration on ethics and data protection in artificial intelligence, Brussels, 23 October 2018.
[230] Ibid.
[231] See https://icdppc.org/public-consultation-ethics-and-data-protection-in-artificial-intelligence-continuing-the-debate/
[232] See, for example: UK House of Lords Artificial Intelligence Committee's report, AI in the UK: ready, willing and able?, published in April 2018

With this conclusion in mind, academics Josh Cowls and Luciano Floridi published, in July 2018, a 'Prolegomena to a White Paper on an Ethical Framework for a Good AI Society'.[233] This work aims to focus on the commonalities and noteworthy differences across the following prominent sets of principles: the Asilomar AI Principles; the Montréal Declaration for Responsible AI; the IEEE principles; the EGE principles; and the principles included in the report 'AI in the UK: ready, willing and able?' developed by the UK House of Lords Artificial Intelligence Committee.

Based on a comparative assessment of the 44 principles detailed across these five reports, Cowls and Floridi found 'an impressive degree of coherence and overlap'[234] and mapped the principles to four core principles commonly used in bioethics (beneficence, non-maleficence, autonomy and justice) and an additional new principle (explicability). The resulting output aims to unite what was previously five sets of principles.

- **Beneficence:** promoting well-being, preserving dignity and sustaining the planet;
- **Non-maleficence:** privacy, security and "capability caution";
- **Autonomy:** the power to decide;
- **Justice:** promoting prosperity and preserving solidarity; and
- **Explicability:** enabling the other principles through intelligibility and accountability.

4.4 Working groups and committees carrying out research and fostering collaboration and an open dialogue

Research and advocacy: FAT/ML, TransAlgo, AlgorithmWatch, algodiv and MIRI

Multiple institutions have ongoing research, advocacy and awareness raising workstreams in the context of AI and algorithmic decision-making. One of the most well-known banners in this space is **FAT/ML**; increasingly used as an umbrella term for a series of academic workshops, often tied to computer science conferences, attempting to bring together a growing community of researchers and practitioners concerned with fairness, accountability and transparency in machine learning systems[235]. In addition to the development of principles and a social impact statement for accountable algorithms (discussed above)[236], a FAT/ML website also provides a list of relevant events[237] and projects involving academia and other research institutions, namely *Explainable Artificial Intelligence (XAI)*[238], and *On algorithmic fairness, discrimination and disparate impact*[239], amongst others[240].

Beyond a well-known (albeit relatively static) resource repository, FAT/ML has been the banner under which annual one-day interdisciplinary workshops have run since 2014. These workshops have been academic venues considering issues of fairness, accountability and transparency largely from the perspective of computer science and optimisation. In 2018, the FAT/ML

[233] Cowls, Josh and Floridi, Luciano, Prolegomena to a White Paper on an Ethical Framework for a Good AI Society (June 19, 2018). Available at SSRN: https://ssrn.com/abstract=3198732 or http://dx.doi.org/10.2139/ssrn.3198732
[234] Ibid.
[235] See http://www.fatml.org/
[236] See http://www.fatml.org/resources/principles-for-accountable-algorithms
[237] For a full list see http://www.fatml.org/resources/relevant-events
[238] See https://www.darpa.mil/program/explainable-artificial-intelligence
[239] See http://fairness.haverford.edu/
[240] For a full list see http://www.fatml.org/resources/relevant-projects

workshop was co-located with the 2018 International Conference on Machine Learning (ICML), which itself has since developed a track of work on fairness in the main conference. Furthermore, from 2014 to 2018 the number of contributions to the workshop has risen from eight presentations to 10 presentations and 20 posters. Similar workshops, such as FATREC (for recommender systems) and Ethics in NLP (for natural language processing) have also begun to complement computer science fields elsewhere. Some of the same individuals as those involved in FAT/ML have established the ACM FAT* conference, which unlike FAT/ML is more formalised initiative attempting to bring together interdisciplinary work in the field, and to act as a publishing venue.

Similarly, **TransAlgo**,[241] provides a collaborative platform for the development of knowledge and a culture for the production, analysis and evaluation of responsible and ethical algorithms and databases. The aim of TransAlgo is to raise awareness and facilitate the adoption of greater transparency. Amongst others, it counts on the support of the French government, the Conseil National du Numérique (CNNum) and the Commission Nationale Informatique & Libertés (CNIL).

In addition to providing updated news and events on algorithmic and data accountability and transparency, TransAlgo also provides users with a scientific repository of resources and tools, which are currently classified across three areas: *Application Areas*, *Social Ethical and Legal Issues*, and *Types of Algorithmic Systems*[242]. TransAlgo also maintains five scientific working groups to discuss the following topics:

- **Search engines and recommending systems:** discussing topics of neutrality, non-discrimination and explicability;
- **Apprenticeship and confidence:** discussing topics of robustness, bias and reproducibility;
- **Confidentiality and consent:** discussing topics of privacy and monitoring information flows;
- **Neutrality and metrology of communication networks:** discussing net neutrality; and
- **Influence, misinformation and impersonation:** discussing how decision-making algorithms are used.

In slight contrast with the two initiatives mentioned above, **AlgorithmWatch**[243], a Berlin-based organisation and a member of the EU HLEG on AI[244], has a more dedicated focus on producing original research and policy recommendations. It describes itself as "a non-profit research and advocacy organisation to evaluate and shed light on algorithmic decision-making processes that have a social relevance"[245]. Specifically, AlgorithmWatch analyses the effects of algorithmic decision-making in society and points out and explains ethical conflicts to the general public, while providing a public networking and engagement platform. It has also published a dedicated *Algorithmic Decision-Making Manifesto*[246], which aims to end the trend of algorithmic

[241] See https://www.transalgo.org
[242] See https://www.transalgo.org/ressources-outils/
[243] See https://algorithmwatch.org/en/
[244] See https://algorithmwatch.org/en/eu-high-level-expert-group-on-artificial-intelligence/
[245] See https://algorithmwatch.org/en/algorithm-watch-mission-statement/
[246] See https://algorithmwatch.org/de/das-adm-manifest-the-adm-manifesto/

decision-making procedures as 'black boxes'. AlgorithmWatch is funded by independent German research foundations, the Hans Böckler Foundation and the Bertelsmann Foundation.

With a more technical focus, **algodiv**[247] and the **Machine Intelligence Research Institute** (MIRI)[248] provide original research on algorithmic recommendation practices and techniques for transparent AI and machine learning approaches, respectively. The main goal of these two projects is to publish original research; however, their outputs are more technically-focussed than the above and therefore more accessible to those who are more literate in the design and development of AI and machine learning systems:

- **Algodiv:** With funding from the French National Research Agency (ANR) in the period 2016-2019, algodiv aims to conduct interdisciplinary research on issues of information diversity in online communities and the effect of algorithms. The project, coordinated by the Centre Marc Bloch, brings together scientific partners from government (the Centre d'Analyse et de Mathématique Sociale, a unit of the Centre National de la Recherche Scientifique – CNRS), academia (the Laboratoire d'Informatique de Paris 6) and industry (Orange Labs).
- **MIRI:** Aims to understand the societal impact of AI technologies, including specific reference to issues of robustness and safety. A US-based non-profit, MIRI has received donations from more than 3,500 donors since its initiation in 2000 (until February 2013, it existed as the Singularity Institute), including from the Thiel Foundation, prominent cryptocurrency developers (including Ethereum and Ripple) and the Open Philanthropy Project.

'A smart web for a more equal future', an initiative by the World Wide Web Foundation

The **World Wide Web Foundation** is currently carrying out a wide range of projects across different themes, but all aggregated under the umbrella of its main mission: *Establishing the open Web as a basic right and a public good*[249]. With specific regard to AI and algorithmic decision-making and following Sir Tim Berners-Lee's call for increased algorithmic transparency[250], the Foundation has launched a white paper series entitled *Opportunities and risks in emerging technologies*[251]. This research effort has focussed on creating and disseminating an understanding of how new technologies are shaping society, where they present opportunities to make people's lives better, and where there are potential risks. The series is comprised of three papers which individually address the topics of AI deployment in low and middle-income countries, the application of the concept of algorithmic accountability to different contexts, and an overview of the personal data ecosystem in low and middle-income countries[252].

Whilst the two papers on AI and personal data provide a comprehensive overview of the benefits and potential risks of emerging technologies in specific geographical scopes, the

[247] See http://algodiv.huma-num.fr/objectives/
[248] https://intelligence.org/research/#RWM
[249] See https://webfoundation.org/our-work/
[250] See https://www.theregister.co.uk/2017/03/12/tim_berners_lee_web_at_28_letter/
[251] See https://webfoundation.org/research/white-paper-series-opportunities-and-risks-in-emerging-technologies/
[252] Ibid.

paper on algorithmic accountability provides a more global account of issues directly associated with algorithmic decision-making. Specifically, the paper notes that *"[…] algorithms[253] have become the backbone of many business models deployed worldwide. In the public sector — particularly in Europe and the US — algorithmic decision-making has emerged alongside broader policy trends of the last decade such as open government and evidence-based decision-making and is now starting to be used in high-stakes areas such as criminal justice"*[254]. Through a review of data collected from interviews with global experts, workshops and content research, the Foundation provides a clear explanation of algorithmic decision-making and the challenges it poses to the understanding of accountability across different contexts[255].

In particular, the publication highlights two key challenges – **algorithmic harm** and **algorithmic discrimination** – and discusses issues, such as fairness, transparency and accountability.

4.5 Policy and technical tools

AI NOW Institute: Algorithmic Impact Assessments and Algorithmic Accountability Policy Toolkit

The **AI NOW Institute**, New York University, is a multi-disciplinary research centre dedicated to understanding the social implications of artificial intelligence. AI NOW receives funding from Microsoft and Google, as well as philanthropic foundations including MacArthur and Ford. The institute focuses on four core domains:

- Rights & Liberties
- Labour & Automation
- Bias & Inclusion
- Safety & Critical Infrastructure.

Through its report "Algorithmic Impact Assessments: a practical framework for public agency accountability"[256], the AI NOW Institute recently proposed a **framework to support public agencies and citizens monitor and understand AI and algorithmic decision-making systems**. Specifically, the framework sets out 5 key elements that public agencies should consider when carrying out an algorithmic impact assessment, namely[257]:

- *Agencies should conduct a **self-assessment of existing and proposed automated decision systems**, evaluating potential impacts on fairness, justice, bias, or other concerns across affected communities;*
- *Agencies should develop meaningful external researcher review processes to discover, measure, or track impacts over time;*
- *Agencies should provide notice to the public disclosing their definition of "automated decision system," existing and proposed systems, and any related self-assessments*

[253] The Foundation provides its own definition of "algorithm", which is as follows: "[…] code and data as well as the greater socio-technical assemblage that includes algorithm, model, target goal, data, training data, application, hardware – and connect it all to a broader social endeavour"
[254] See https://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf
[255] The topic of algorithmic accountability is also further explored from a policymaking perspective in the Center for Data Innovation's recent paper "How Policymakers can foster algorithmic accountability", which can be accessed at http://www2.datainnovation.org/2018-algorithmic-accountability.pdf
[256] See https://ainowinstitute.org/aiareport2018.pdf
[257] Ibid.

and researcher review processes before the system has been acquired;
- *Agencies should solicit public comments to clarify concerns and answer outstanding questions;*
- *Governments should provide enhanced due process mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased, or otherwise harmful system uses that agencies have failed to mitigate or correct.*

The AI NOW Institute has also stated that algorithmic impact assessments cannot be the single solution in solving potential societal problems brought about by decision-making systems, but that they do provide an important mechanism to engage and inform the public, policymakers and researchers.

Building on its work on Algorithmic Impact Assessments, the AI NOW Institute, in October 2018, published an **Algorithmic Accountability Policy Toolkit**.[258] This Toolkit demonstrates and explains the types of algorithmic systems used within government, including in: Human Resources; Public Health; Criminal Justice; and Education. It also presents a list of relevant literature.

UnBias: emancipating users against algorithmic biases for a trusted digital economy

UnBias[259] is a UK Engineering and Physical Sciences Research Council (EPSRC) research project which aims to provide ethical guidelines and policy recommendations with the objective of ensuring trust and transparency in the internet. Furthermore, UnBias has continuously been developing a "Fairness Toolkit",[260] which aims to promote awareness and stimulate a public dialogue about how algorithms shape online experiences, as well as to generate reflections on possible changes to address issues of online unfairness. The toolkit was developed with the input of young people and other key stakeholders and contains:

i. a handbook, which provides an overview of the different components in the toolkit;
ii. 63 "awareness cards", which are designed to build awareness of how bias and unfairness can occur in algorithmic systems;
iii. TrustScapes, which contain visualisations of issues with algorithmic bias, data protection and online safety, as well as potential solutions to ensure a safe and fair online environment;
iv. MetaMaps, which comprise posters for stakeholders in the technology industry, policymaking, public sector organisations and academia to respond to the TrustScapes generated by users; and
v. value perception worksheets, which provide an assessment framework of the toolkit itself[261].

The Center for Democracy & Technology's (CDT) 'Digital Decisions Tool'

[258] AI NOW, Algorithmic Accountability Policy Toolkit, October 2018. https://ainowinstitute.org/aap-toolkit.pdf
[259] See https://unbias.wp.horizon.ac.uk/our-mission/
[260] See https://unbias.wp.horizon.ac.uk/fairness-toolkit/
[261] See https://unbias.wp.horizon.ac.uk/fairness-toolkit/

The **CDT** is currently carrying out a project entitled "digital decisions". Its main goal is to support engineers and product managers of decision-making algorithms in mitigating against the risks of designing unfair, discriminatory and harmful systems[262]. As part of the project, CDT has created an **interactive "digital decisions tool"** which works by translating principles for fair and ethical automated decision-making into a series of questions that can be addressed during the process of designing and implementing an algorithm[263]. Specifically, these questions enquire about the choices of developers, such as training and input data, testing methodologies, as well as processes and checks for assessing risk and ensuring fairness. The CDT has stated that this tool was informed by extensive research[264], and that it is currently being subject to continuous iteration, testing and re-evaluations.

Accenture's and Alan Turing Institute's algorithmic fairness evaluation tool

In recognising that autonomous systems can be responsible for life-changing decisions about employment, finance, and many other areas, **Accenture**, in partnership with the **Alan Turing Institute**, has recently developed an algorithmic "Fairness Tool"[265,266,267]. Complementary to the UnBias toolkit and the CDT's digital decisions tool, this "Fairness Tool" aims to **scrutinise the data** that goes into an algorithm through a process of identification and removal of any coordinated influence between sensitive variables (e.g. race, gender) that may lead to an unfair outcome. Furthermore, the tool is able to ensure that the rates of false positives and negatives of any given group in a data set are "fairly distributed"[268]. When publishing the tool, Accenture also acknowledged that introducing fairness in an algorithmic system might decrease its accuracy, and that the tools is able to display the extent to which that may have happened.

Data Transparency Lab's Technical Programme

The **Data Transparency Lab** (DTL) is an inter-institutional collaboration, seeking to create a global community of technologists, researchers, policymakers and industry representatives working to advance online personal data transparency through scientific research, innovation and design[269]. Whilst not explicitly involved in the algorithmic decision-making debate, the DTL aims to connect key stakeholders (i.e. researchers, developers, policymakers and industry players) to co-develop solutions that ensure data transparency and provide technical infrastructure and support to foster a "healthy data sharing ecosystem".

Furthermore, since 2015, the DTL has provided grants for tools that support the development of software on data privacy and transparency across a range of areas, including algorithmic bias and discrimination; PII leakage; reverse engineering online pricing; transparency of AI an ML algorithms; and explainable AI.

---

[262] https://cdt.org/blog/digital-decisions-tool/
[263] https://cdt.info/ddtool/
[264] https://cdt.org/issue/privacy-data/digital-decisions/
[265] See https://www.accenture.com/gb-en/blogs/blogs-cogx-tackling-challenge-ethics-ai
[266] See https://www.turing.ac.uk/collaborate-turing/data-study-groups/accenture-challenge-fairness-algorithmic-decision-making
[267] See https://www.bloomberg.com/news/articles/2018-06-13/accenture-unveils-tool-to-help-companies-insure-their-ai-is-fair
[268] See https://www.accenture.com/gb-en/blogs/blogs-cogx-tackling-challenge-ethics-ai
[269] See https://datatransparencylab.org/about/

A number of relevant tools have already been developed by the DTL and are currently available, including:

- **FA*IR:** aimed at ensuring fair rankings in search engines;
- **AdAnalyst:** aimed at increasing transparency in the Facebook advertising ecosystem;
- **Lumen:** aimed at informing and enabling users to control the communications of their mobile apps with tracking systems;
- **$heriff:** aimed at enabling users to search for traces of price discrimination in online platforms; and
- **ReCon:** aimed at informing users of the information sent by their mobile applications to third parties.

EthicsToolkit.ai – an Ethics & Algorithms Toolkit

In similar fashion to the framework for algorithmic impact assessments published by the AI NOW Institute, the **Ethics & Algorithms Toolkit** focusses on empowering and supporting governments and public agencies to understand and mitigate the risks of algorithmic decision-making[270]. Co-developed by GovEx, the City and County of San Francisco, Harvard DataSmart, and Data Community DC[271], the toolkit is meant to provide cities with a comprehensive understanding of the implications of using an algorithm, while clearly stating its potential risks and mitigation measures. The toolkit itself is made up of two main parts: **algorithm assessment**, and **algorithm risk management**. The algorithm assessment part takes the user of the toolkit through a process of understanding and assessing the impact of the algorithm in question, as well as the risks of appropriate data use, accountability, methodology, and historic and technical biases. The algorithm risk management part provides a list of potential mitigation maps and an accompanying risk-to-mitigation measure map, which allows users to understand which measures should be applied to specific characteristics identified when scoping and assessing the algorithm[272].

4.6 Intermediate findings

There is **significant effort being directed towards tackling the challenges facing algorithmic decision-making by industry, civil society, academia and other interested parties**. This is true across all categories of initiatives examined and relates to all of the perspectives discussed above. In particular, there are a large number of initiatives aimed at promoting responsible decision-making algorithms through codes of conduct, ethical principles or ethical frameworks.

Including this type of initiative, we have clustered the initiatives identified in four main types. Here, we summarise the findings on the work being conducted in each category, noting, in

[270] See http://ethicstoolkit.ai/
[271] The motivation for the co-development of this tool is outlined in a research paper published by DataSF, which can be accessed at https://docs.google.com/document/d/1ivEbJwiP8MiuPOJJEsZG2xn6fR2y3OnxdAtx8yb3IN0/preview
[272] Both parts of the toolkit are available, respectively, at:
https://drive.google.com/file/d/1JKs8jUlNt5lBHr_2hs6Gv-8VS5kgvohf/view
https://drive.google.com/file/d/1dP2VznxErwlzQq1DfpDjbPxEFPbu5Q0K/view

particular, that data on the impact and take-up of these initiatives is not available in the majority of cases:

Standardisation efforts: ISO and the IEEE are two of the most prominent global standards bodies, with the buy-in and cooperation of a significant number of national standards bodies. As such, it is important that these organisations are working towards tackling a number of these challenges.

The final effort documented here, outside of the scope of the ISO and the IEEE, is the Chinese White Paper on Standardisation. Although no concrete work has been conducted, this document illustrates that stakeholders currently involved in the standardisation process in China – a multi-disciplinary group – are considering algorithmic decision-making from all the key perspectives being discussed.

Codes of conduct, ethical principles and frameworks: As mentioned above, there are a vast number of attempts to govern the ethics of AI development and use with no clear understanding or reporting on take-up or impact. These initiatives have been initiated by stakeholders from all relevant groups, in some cases in isolation but also through multi-disciplinary efforts. Furthermore, much of this work attempts to tackle the challenges facing algorithmic decision-making from multiple perspectives. For instance, the ethical principles developed by the Software and Information Industry Association (SIIA) explicitly discuss the need for transparency and accountability; and the Asilomar Principles, which cover, in particular, topics of fairness, transparency, accountability, robustness and privacy.

Interesting work that stands out and could be beneficial on a higher plane includes the work of Algorithmenethik on determining the success factors for a professional ethics code and the work of academics Cowls and Floridi, who recognised the emergence of numerous codes with similar principles and conducted an analysis across some of the most prominent examples. Cowls and Floridi's work is also valuable as it ties the industry of AI development and algorithmic decision-making to long established ethical principles from bioethics. The elements of learning these examples bring from established sectors can be extremely useful.

Working groups and committees: The initiatives examined have primarily been initiated by civil society organisations (including, for example, AlgorithmWatch and the Machine Intelligence Research Institute) with the aim of bringing together a wide variety of stakeholders. Outputs of these initiatives tend to include collaborative events, such as the FAT/ML workshops, or research papers and advice, such as the World Wide Web Foundation's white paper series on *Opportunities and risks in emerging technologies*. As for the above, this type of initiative is often focused on tackling the challenges facing algorithmic decision-making from multiple perspectives. For instance, AlgorithmWatch maintains scientific working groups, which, in the context of various challenges, discuss, amongst others, topics of non-discrimination and bias, privacy and algorithmic robustness. Furthermore, no clear information on the impact of these initiatives is currently available.

Policy and technical tools: In this category, the initiatives examined have been developed by academic research groups (e.g. the work of NYU's AI Now Institute and the UnBias research project), civil society (e.g. the Digital Decisions Tool of the Center for Democracy and Technology) or multi-disciplinary groups (e.g. the EthicsToolkit.ai developed through collaboration between academia and policy-makers). In terms of how these tools address the challenges facing algorithmic decision-making, they tend to focus on specific challenges; a clear example being the 'Fairness Toolkit', developed by the UnBias research project.

5. Index of policy initiatives and approaches

This section provides an overview of the existing policy approaches by distinguishing between:

- EU wide approaches,
- EU national approaches, and
- Approaches by third countries.

This report does not provide an exhaustive and systematic account of algorithmic decision-making and AI approaches of all EU Member States. Instead, it focusses on specific EU and third countries[273]. For the EU countries, the AI landscape in Ireland, UK, France, Netherlands, Italy, Poland, Estonia, Spain, Denmark, Finland and Germany has been assessed. In third countries, Canada, the USA, South Korea, China, Japan, India and Singapore have been scrutinised. This section will also take a closer look at the efforts regarding AI of the separate United Nations agencies.

Table 2 summarises the approaches by highlighting, for each of the three geographical areas:

- the name and timeline of the initiative;
- the geographical scope (depending on whether the initiative applies on a federal level or not);
- the type of initiative (such as legislations, bills, policy documents, etc.);
- the material scope of the initiative (e.g. what aspects of AI and algorithmic decision-making does it cover, which sector(s) are covered);
- an analysis of whether the link to algorithmic decision-making is implicit or explicit;
- what the expected impact is (what results can be expected from the law or policy document); and
- the involved stakeholders (e.g. government, industry, society as a whole, academia, etc.).

Following this table, the section provides a more detailed account of the different policy approaches and options retrieved on each country.

Where appropriate, this section points out opportunities policies and policy initiatives might provide for other countries or institutions. While some of the policies might be directed at very specific problems of their respective countries or organisations (e.g. military aspects of AI for the US or NATO countries), other initiatives or pieces of legislation might be a reference to tackle issues in many countries or even on the EU-level (e.g. the Estonian Kratt-Law, which defines a clear legal status for AI-operated machines). Synergies across countries, for example

[273] In the current version of this report, and with regard to national-level policy measures within the EU, this section presents the current algorithmic decision-making policy landscape and ongoing policy debates in the UK, Germany, France, the Netherlands, Poland, Ireland, Italy, Estonia, Spain, Denmark and Finland. The first 4 countries were selected given their advancement in the field of AI as reported in the media, as well as the frequency of debate in the field at a policy-making level, as evidenced by the publication of government position papers or calls for regulation / legislation. Ireland were chosen given their level of digital maturity, as well as the representation of the digital technology commercial sector in those countries. Poland and Italy were selected given the recent cases of 'unfair' profiling of unemployed citizens and dissemination of 'fake news' through social media-based advertisement networks, respectively. Estonia introduced a novel way to look at AI as legal entities, while Spain recently appointed an expert group to develop a white book on artificial intelligence and Big data. Finally, the Denmark and Finland not only work together in the Nordic Council of Ministers, but also both released national strategies on how to respond to and use AI in the near future.
The analysis presented herein is thus to be regarded as a work in progress, for which the geographical scope will be expanded in future iterations of the report.

regarding the combination and internationalisation of national AI strategies, might lead to a net welfare gain in terms of profitability, applicability and efficiency.[274]

To set the policy initiatives into context, this section also makes references to the previous sections with regard to the six areas of discussion facing algorithmic decision-making: Fairness & Equity; Transparency & Scrutiny; Accountability; Robustness & Resilience; Privacy as well as Liability. This should give the reader a clearer understanding of which areas are tackled on which level, which topics are in the center of the debate in certain countries and what emphasis is set by the various stakeholders involved in the process. Extracting the core messages of certain policies also facilitates the peer-review process of this document, as it makes it easier to specifically "target" certain sections to academics and experts in the respective fields. In addition, a clear thematic framework facilitates the formulation of questions for interviews of stakeholders involved in the policy processes.

In conclusion, this section is aimed at giving policy-makers the opportunity to compare national and international efforts and implement or create, where needed, similar measures or actions. It also makes it easier to distribute certain sections in the peer-review process, which is crucial for the validation of the core findings.

At present, this section reflects the results of an extensive desk-based research exercise. However, in the next stage of the project, we will undertake two key research tasks to increase the relevance, utility and effectiveness of this report:

- conduct a peer-review process on the report; and
- undertake an extensive interview schedule with stakeholders covering all interested groups, including academia, industry, public sector and civil society.

In particular, in relation to this section, these additional research activities will focus on: i) validating the content of this report; and ii) providing further information on why these policy initiatives were implemented, how they work theoretically and in practice and their expected or actual impact.

[274] As could be expected for the European AI Alliance.

Table 2: Index of policy initiatives and approaches

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
European Union	**General Data Protection Regulation (GDPR)[278]** (Adopted in 2016, in effect since 2018)	**EU-wide application**	EU Regulation	**It covers the processing of personal data. Article 22 specifies that decisions based solely on automated processing, which produces legal effects are not permitted unless certain requirements are fulfilled. Art 13, 14, 15 include further provisions on the right to meaningful information about the existence of automated decision-making, including profiling.**	**The impact is likely to be the increased transparency and protection from automated decision-making. It also provides a clearer framework under which circumstances its application is possible.[279]**	**The regulation was adopted by the European Parliament (EP) and the Council of the European Union**	**Stakeholders providing and processing data in the EU (individuals, government, industry, NGOs, etc.)**

[275] The main sources can be found in the footnotes next to the respective policy initiative names and, where appropriate, in the specific boxes they are concerned with.
[276] Geographical coverage / scope
[277] The expected impact provides an explanation of the primary intent of a policy in primary literature and/or experts' opinions in secondary literature.
[278] 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)
[279] See e.g. https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-else-do-we-need-to-consider-if-article-22-applies/

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
	Promoting fairness and transparency for business users of online intermediation services[280] (Proposal of the European Commission in 2018; currently under discussion of the European Parliament and Council))		EU Regulation	**It imposes transparency obligations on online platforms with regards to the main parameters used for ranking offers from their business users or ranking webpages on search engines.**	**The proposal is aimed at establishing trust in trading practices between online platforms and their business users and search engines and corporate websites.**[281]	**The proposal was designed by the European Commission**	**Mainly affects online platforms (online intermediation service providers and search engine) and operators of businesses online, with potential impacts on consumers**
	Markets in Financial Instruments Directive (MiFID II)[282] (approved in 2014, in effect since 2017)		EU Directive	**It deals with the use of algorithms in the financial markets and regulates algorithmic trading. It specifies that investment firms are obliged to provide competent authorities with a description of the nature of their**	**The purpose is to increase transparency of algorithmic trading and to introduce risk control systems against harmful decisions taken by algorithms.**	**Passed by the European Parliament and the Council. The implementation process and ongoing operation is monitored by the European Securities and Markets Authority (ESMA)**[285] **National law transposes the**	**Primarily targeted at firms engaged in trading in the financial sector, like investment firms as well as trading venues.**[286]

[280] European Commission (2018) Proposal for a Regulation of the European Parliament and of the Council on promoting fairness and transparency for business users of online intermediation services
[281] Ibid.
[282] Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments amending Directive 2002/92/EC and Directive 2011/61/EU
[285] See https://www.esma.europa.eu/policy-rules/mifid-ii-and-mifir
[286] See http://deutsche-boerse.com/dbg-en/regulation/regulatory-dossiers/mifid-mifir/mifid-i-to-mifid-ii/market-structure/algo-hft

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				algorithmic trading strategies.[283] Also it specifies that trading venues[284] are taking proficient steps to prevent distortions caused by algorithmic trading.		**Directive in each Member State.**	
	"Artificial Intelligence for Europe"[287] (Introduced in 2018)		EU Commission Communication	**Launches a series of initiatives to boost EU competitiveness in AI, as well as specific measures targeted at increasing transparency for consumers of algorithmic decision-making (such as the development of safety frameworks and guidelines on cases where algorithmic decision-making affects**	**Contribute to the wide uptake of ethical principles in developing AI.[288]**	**Communications from the European Commission**	**Industry developing and deploying AI, citizens.**

[283] See http://deutsche-boerse.com/dbg-en/regulation/regulatory-dossiers/mifid-mifir/mifid-i-to-mifid-ii/market-structure/algo-hft
[284] Which are alternatives to traditional stock exchanges.
[287] European Commission (2018) 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, SWD(2018) 137 FINAL
[288] Ibid.

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				consumers) It also includes a section on ethical and human rights considerations and announces actions related to this.			
	Measures to effectively tackle illegal content online[289] (Proposed in 2018)		EC Recommendation	**Incentivise hosting service providers to implement measures to proactively detect, identify and remove or disable content that propagates or incites to terrorism**	**Clearer framework on how notices of illegal content should be processed. Closer operations of online platforms with Member states and among themselves to tackle illegal content. Faster and more effective law enforcement notification regarding terroristic threats.[290]**	**Formulated by the European Commission**	**The execution of punishment in case of non-compliance would be done by the respective bodies of law enforcement. The measures would mainly affect businesses operating online.**
	Digital Single Market Opportunities and Challenges for Europe[291]		EC Communication	**Acknowledges online platforms as a driver of innovation and growth in the**	**Support self-regulation and co-regulation in order to foster strong platform**	**Formulated by the European Commission.**	**Mainly affects providers of online platforms.**

[289] European Commission (2018) Recommendation on measures to effectively tackle illegal content online. p.1-2
[290] European Commission (2018) Fact Sheet on the Commission Recommendation on measures to effectively tackle illegal content online
[291] European Commission COM(2016) 288 final (2016), Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Online Platforms and the Digital Single Market Opportunities and Challenges for Europe

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
	(Introduced in 2016)			**Digital Single Market. Aims at effectively simulating innovation, while protecting legitimate interests of consumers and other users**	**ecosystems in Europe. Strengthen the industry's compliance with applicable EU law.[292]**		
Estonia	**Personal Data Protection Act[293]** (entered into force in 2008 and about to be updated)	**Estonia**	Legislation	**Article 17 includes the prohibition of the use of unsupervised "automated decision" making in certain cases, especially in case of a lack of disclosure or a possibility to object.**	**Similar to Article 22 of the GDPR, Article 17 clearly specifies under which circumstances an automated process can make legally binding decisions for or with a person.[294]**	**Passed by the national parliament.**	**The law affects Individuals, government, industries, NGOs, etc.**
	Algorithmic-liability law, also known as Kratt-Law[295] (In discussion)	**Estonia**	Legislation	**A bill which allows the real-life application of fully autonomous information systems, a clear legal framework regarding their**	**Clarification about responsibility and liability issues regarding decisions taken by Artificial Intelligences.[296]**	**State authorities, universities, companies and independent experts are involved in the law-making process.**	

[292] Ibid.
[293] See https://www.riigiteataja.ee/en/eli/512112013011/consolide
[294] Ibid, as well as the Regulation (EU) 2016/679 of the European Parliament and of the Council (GDPR)
[295] See https://e-estonia.com/ai-and-the-kratt-momentum/
[296] Ibid.

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				decisions as well as a sufficient concept of supervision			
France	**Digital Republic Act 2016**[297] (Updated by decree in 2017)	**France**	Legislation	**The legislation places requirements on government bodies that make decisions solely or partially by algorithmic systems to give individuals certain transparency rights regarding which decisions were made.**	**The law aims at increasing transparency in cases where public administrations take automated decisions on citizens by requiring administrations to provide information on the parameters and data used for taking those automated decisions.**[298]	**The legislation was prolonged by Presidential decree.**	**Art. L. 311-3-1 and art. L. 312-1-3 of the- law are primarily targeted at public administrations.**
	AI for Humanity Strategy[299] (Initiated in 2018)	**France**	Policy Document	**The document establishes a strategy to turn France into a global leader in AI research, training, and industry. Among others it includes the need to**	**The expected impact of the policy document is to place France at the forefront of AI whilst minding that algorithmic decision-making is transparent and**	**The strategy was initiated by President Macron, who commissioned Cédric Villani to lead the project.**	**It affects individuals, government, industry, NGOs and other parts of society.**

[297] Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.
[298] Ibid.
[299] République Française – Le Gouvernement (2018) *Rapport de synthèse – France Intelligence Artificielle. See for example:* Available at https://www.gouvernement.fr/en/artificial-intelligence-making-france-a-leade

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				develop an ethical framework for AI, in order to ensure that its development and use is transparent, explainable, and non-discriminatory.	**not discriminatory.[300]**		
	New Personal Data Protection Act[301] (In action since 2018)	**France**	Legislation	**The legislation aims at integrating the provisions of the GDPR into French legislation. Article 21 discusses exceptions for the application of solely algorithm-based legal decisions and the rights natural persons have to observe and object to those decisions.**	**Regarding automated decision-making, the law provides exceptions to allow for automated decision-making without human intervention, while outlining possibilities to object these decisions.[302]**	**The legislation was passed by the National Assembly and the Senate.**	**individuals, the government, industry, NGOs and other parts of society**
	French Digital Council[303] (Founded in 2011)	**France**	Agency	**The French Digital Council is an institution set up by the**	**The aim of the Council is to advise the government on**	**Composed of 30 voluntary experts representing the digital economy,**	**It affects individuals, government, industry, NGOs**

[300] See https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf
[301] See See https://blogdroiteuropeen.files.wordpress.com/2018/06/olivia.pdf
[302] Ibid.
[303] See French presidential decree n°2011-476.

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				government composed of 30 experts.	**issues relating to algorithmic decision-making and ethical considerations thereof.**[304]	**academia and civil society.**	**and other parts of society**
Germany	**White Paper on AI**[305] (Issued in 2018)	**Germany**	Policy Document	**The German Government has published the White paper focusing mainly on how to increase the uptake of AI. However, the paper also addresses the need to establish a framework on transparency and auditability of algorithms.**	**The likely impact could be the introduction of a series of oversight and control mechanisms on algorithms at a later stage.**[306]	**German Federal Government**	**Not clear yet, but It could affect individuals, the government, industry and NGOs, etc.**
	Enquete Commission[307] (Approved in 2018)		Expert Group	**The Enquete Commission of the German Bundestag investigates ethical and societal aspects of algorithmic decision- making and will present a**	**The Commission's investigations and recommendations for action will inform the policy debate and may lead to more concrete suggestions on**	**The commission includes Members of Parliament as well as AI experts, such as Dr Stefan Heumann of the Technology think tank *Stiftung Neue***	**See above.**

[304] *Conseil National du Numérique* (CNNum), whose mission and objectives are listed at https://cnnumerique.fr/missions/
[305] See https://www.bmwi.de/Redaktion/DE/Downloads/E/eckpunktepapier-ki.pdf?__blob=publicationFile&v=4
[306] Assessment based on the source above.
[307] See https://www.bundestag.de/dokumente/textarchiv/2018/kw26-de-enquete-kommission-kuenstliche-intelligenz/560330

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				final report on potential legislative initiatives by the end of 2020 to the Parliament.	**ethical issues of algorithmic decision-making.[308]**	***Verantwortung* and Prof Jörg Müller-Lietzkow of Paderborn University.[309]**	
Italy	**Artificial Intelligence: At the Service of Citizens[310]** (Issued in 2018)	**Italy**	Policy document	**The Italian AI strategy outlines the risks of algorithmic decision-making and the aspects that need to be borne in mind in the governance debate such as data quality and neutrality, responsibility, transparency and privacy.**	**Highlighting these aspects could result into discussions on a regulatory framework on the use of algorithms in decision-making processes.[311]**	**Edited by the AI Task Force (see below) and affecting individuals, government, industry, NGOs, etc.**	
	AI Task Force[312] (Launched in 2017)		Agency	**The Agency for Digital Italy (AGID) has established an AI tasks force which aims to among others explore**	**More practical examples on the decision-making through algorithms will inform the debate.[313]**	**Issued by the government, execution involves several stakeholders from the government to academia,**	**Individuals, the government, industry and NGOs and a multitude of other parts of society.**

[308] Ibid.
[309] See e.g. https://www.bundestag.de/ausschuesse/weitere_gremien/enquete_ki#url=L2Rva3VtZW50ZS90ZXh0YXJjaGl2LzIwMTgva3c0NS1wYS1lbnF1ZXRlLWtpLzU3NTcxMg==&mod=mod569768
[310] See https://ai-white-paper.readthedocs.io/en/latest/, see also: https://ia.italia.it/en/task-force/
[311] Ibid., Challenge 1
[312] See https://futureoflife.org/ai-policy-italy/
[313] See https://libro-bianco-ia.readthedocs.io/en/latest/

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				the ethical and social implications of the use of AI in decision-making processes.		**including 30 direct members and around 450 community members from many sectors.**	
Poland	**Visegrád 4 countries' thoughts on the Artificial Intelligence and maximising its benefits ahead of release of the European Commission's Communication on the topic.**[314] (Launched in 2018[315])	**Visegrád countries**	Position Paper	**Deriving from an initiative from Poland, the paper highlights the need to mind the legal, social and ethical aspects of AI and the need to consider how AI can support public administration in decision-making**	**Providing overall strategic guidance on how to approach AI on a policy level.**[316]	**Initiated by Poland and presented at the Visegrád 4 Conference on Artificial Intelligence.**[317]	**It discusses, inter alia, implications of AI for individuals, government, industry, NGOs, etc.**
Spain	**Group of Experts and "Sages" commissioned by the government to develop a White Book on Artificial Intelligence and Big Data**[318] (Launched in 2018)	**Spain**	White Book	**An ethics code which will study the social, juridical and ethical implications that come with the use of new technologies.**	**Provide a framework as well as define limits to work with Artificial Intelligence and Big Data in a legal and ethical way.** [319]	**Developed by a group of experts from government, industry, NGOs and academia.**	**Presumably many parts of society, as a very broad area is covered within the scope of the white book.**

[314] See http://digiczech.eu/wp-content/uploads/2018/04/V4_NON_PAPER_ON_AI_09_04_2018.pdf
[315] See https://www.ideal-ist.eu/event/visegrad-4-conference-artificial-intelligence
[316] Ibid.
[317] See https://www.ideal-ist.eu/event/visegrad-4-conference-artificial-intelligence
[318] See https://www.efefuturo.com/noticia/lado-oscuro-los-algoritmos/
[319] See https://blogthinkbig.com/libro-blanco-inteligencia-artificial

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
United Kingdom	**Centre for Data Ethics and Innovation**[320] (To be launched)[321]	**UK**	Agency	**The Centre is a public body attached to the Department for Digital, Culture, Media and Sport (DCMS). While the precise role of this advisory body will be shaped in the coming years through a consultation, it exists to advise the government to deal with the novel ethical issues raised by algorithmic decision-making.**	**With the help of the Centre, the government will be able to agree on best practices around algorithms, identifying potential new regulations, and setting out measures needed to build trust and enable innovation.**[322]	**Initiated by the government, the centre will be operated by a wider spectrum of stakeholders and experts**	**The results of the analyses of the centre should provide advice to the UK Government on policies and regulations, which could affect a broader spectrum of society.**
	The Government Office for AI[323] (To be launched)		Government department co-ed by DCMS and BEIS	**The core aim of the department is to help the UK to lead in AI in an ethical way by supporting people to develop**	**The aim of the department is to understand existing needs and implement**	**Initiated by the government, executed by several stakeholders**	**Mainly is targeted to make starting a digital business attractive and promoting the adoption of AI by stakeholders in**

[320] See Department for Digital, Culture, Media and Sport (2018) Consultation on the Centre for Data Ethics and Innovation. HM Government, London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/715760/CDEI_consultation_1_.pdf

[321] See https://www.parliament.uk/documents/commons-committees/science-technology/Correspondence/180710-Chair-to-Jeremy-Wright-centre-for-data-ethics-and-innovation.pdf

[322] Ibid.

[323] Department for Digital, Culture, Media & Sport (2018) Press release: World-leading expert Demis Hassabis to advise new Government Office for Artificial Intelligence. HM Government, London. https://www.gov.uk/government/news/world-leading-expert-demis-hassabis-to-advise-new-government-office-for-artificial-intelligence

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				skills for understanding and using AI.	**respective policy initiatives.**[324]		**various sectors of the economy.**
Nordic Countries (Denmark and Finland)	**Strategy for Digital Growth**[325] (Issued in 2018, running until 2025)	**Denmark**	Government strategy	**Allocate 1bn DKK (about €134m) for 38 initiatives of Danish companies to exploit new technologies and generate growth and more wealth for all Danes.**	**The strategy wants to bring Denmark in front of the digital development, while also creating growth and wealth among the Danish population.**[326]	**Initiated and financed by the government**	**The strategy mainly aims to strengthen businesses.**
	Finland's Age of Artificial Intelligence[327] (Issued in 2017, final report to be completed in 2019)	**Finland**	Report	**Includes eight proposals to achieve a successful age of artificial intelligence for Finland. These include provisions to speed up and simplify the adoption of AI, the right application of data as well as strategies to**	**Integrate AI as an active part of every Finn's daily life. Making use of AI in all areas of society, e.g. in health care or manufacturing, ethically and openly.**[328]	**Initiated by the government, drafted by members if academia and experts.**	

[324] Ibid.
[325] Ministry of Industry, Business and Financial Affairs (2018) Press release: New Strategy to make Denmark the New Digital Frontrunner, https://eng.em.dk/news/2018/januar/new-strategy-to-make-denmark-the-new-digital-frontrunner/
[326] Ibid.
[327] Ministry of Economic Affairs and Employment (2017) Publication: Finland's Age of Artificial Intelligence, FI Government, Helsinki. http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf?sequence=1&isAllowed=y
[328] Ibid.

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				attract experts in the field			
	Work in the age of artificial intelligence[329] (Issued in 2018)		Report	**Preparation of a lifelong-leaning reform where every person of working age would be given a skills account or voucher that they could use to update their skills in accordance to workplace and societal changes due to AI developments.**	**The introduction of the voucher and more responsibility for employees, employers and society could create a demand-based market for education and training.**[330]	**Initiated and formulated by the Finnish government.**	**The initiative is targeted at employers and employees equally.**
Australia	**Inquiry on the impact of digital platforms on competition in media and advertising markets.**[331] (Issue paper released in February 2018, final paper expected to be released in 2019)	**Australia**	Report and consultation	**Inquiry and public consultation of the power of digital platforms (e.g. Google, Facebook) on the content of news creators. The report also discusses the degree of transparency of**	**Potential changes in competition policy and better understanding of the influence of digital platforms on news distribution.**[332]	**Initiated by the Australian Competition and Consumer Commission (ACCC) and carried out as a public inquiry.**	**Mainly concerns news creators and digital platforms.**

[329] Ministry of Economic Affairs and Employment (2018) Publication: Work in the age of artificial intelligence, FI Government, Helsinki. http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160980/TEMjul_21_2018_Work_in_the_age.pdf
[330] Ibid.
[331] Australian Competition & Consumer Commission (2018) Media Release: ACCC Seeking views on news and digital platforms inquiry. https://www.accc.gov.au/media-release/accc-seeking-views-on-news-and-digital-platforms-inquiry
[332] Australian Competition & Consumer Commission (2018) Issues paper: Digital platforms inquiry. See https://www.accc.gov.au/focus-areas/inquiries/digital-platforms-inquiry/issues-paper

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				changes in algorithms affecting the visibility of news content on the respective platforms.			
Canada	**Pan-Canadian Artificial Intelligence Strategy**[333] (Announced in 2017)	**Canada**	AI Strategy	**A $125 project which aims at supporting scientific research and the transformation of achievements in AI technology to a broader spectrum of the private and public sector.**	**Improving Canada's position in AI research and training, increasing the productivity in AI academic research and fostering collaboration. Attracting and retaining talent.**[334]	**Launched by the Canadian government and executed by the research institute CIFAR.**	**Mainly aims at supporting the research sector.**
China	**Next Generation AI Development Plan**[335] (Issued in 2017)	**China**	Policy document	**The strategy is mainly about investment in and roll-out of AI but also acknowledges the need to establish laws, regulation and ethical norms surrounding the use of AI. In particular, it is thus necessary to**	**The strategy mentions that research shall be carried out on setting up a framework on ethics and morals as well as a behaviour code for AI product research designers. Furthermore, the**	**Issued by the government**	**Aimed at encouraging an exchange between international stakeholders**

[333] See https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy
[334] Ibid.
[335] China Science & Technology (2017) Newsletter: Next Generation Artificial Intelligence Development Plan Issued by the State Council. Department of International Cooperation Ministry of Science and Technology, P.R. China, Beijing, http://www.chinaembassy-fi.org/eng/kxjs/P020171025789108009001.pdf

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				establish traceability and accountability systems.	**strategy encourages the exchange between stakeholders on an international level.**[336]		
India	**National AI Strategy**[337] (Issued in 2018)	**India**	Policy Document	**The strategy highlights all aspects of AI including among others growth, employment, health and the ethical dimension of AI. On the latter point the strategy highlights the need to conduct research and develop methods to ensure the fairness and transparency of algorithmic decision-making.**	**The AI strategy recommends the government to set up a consortium of Ethics Councils, research centres and the private sector to define the standard practices and monitor their adoption.**[338]	**Initiated by the government, executed by a think tank**	**Focused on research centres and the private sector.**
Japan	**AI guidelines on the assessment of risks of AI**[339] (Issued in 2017)	**Japan & international**	Draft AI Guidelines	**The guidelines establish a number of principles that shall be respected**	**The aim of the draft guidelines is to discuss them in international fora such as the G7 or**	**Issued by the government**	**Aimed at a broad audience in order to achieve a human-centred society**

[336] Ibid.
[337] See http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf
[338] Ibid, p.7.
[339] Prime Minister of Japan and His Cabinet (2016) New release: Public-Private Dialogue towards Investment for the Future, Tokyo, https://japan.kantei.go.jp/97_abe/actions/201604/12article6.html

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				in relation to AI. Most importantly it established the following four principles which are relevant for algorithmic decision-making: The principle of ethics (developers shall take necessary measures so as not to cause unfair discrimination resulting from prejudice included in the learning data of the AI systems) The principle of transparency; Principle of controllability, Principle of privacy.	**the OECD in order to find international consensus on the principles that shall be minded in relation to AI and artificial intelligence.[340]**		
USA	**National Artificial Intelligence Research and Development Strategic Plan[341]** (Issued in 2016)	**US**	Policy Document	**The plan sets out a series of objectives for AI research for academia, industry and**	**The strategy will potentially lead to further research being carried out on the ethical dimension**	**Issued by the government.**	**Targeted at academia, industry and the government itself.**

[340] See http://events.science-japon.org/dlai17/doc/MIC%20-%20France-Japan%20Symposium%2020171025.pdf

[341] National Science and Technology Council (2016) The National Artificial Intelligence Research and Development Strategic Plan. Available at https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				within government. The strategy addresses algorithmic decision-making processes by stating that both data used and the design of AI itself need to be researched to ensure that decisions are fair and transparent.	**of AI and algorithmic decision-making.[342]**		
	Algorithmic accountability bill[343] (Passed in 2017)	**New York**	Law	**The bill is targeted mainly at the New York City agencies that use algorithms in decision-making processes. It establishes a task force with an advisory role on specific measures to ensure that administrations make the use of algorithms transparent and that sufficient safeguard**	**Ensuring the fairness and validity of algorithms used by municipal agencies. [344]**	**Issued by the government of New York.**	**Targeted at public administration.**

[342] See https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf
[343] See https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0
[344] Ibid.

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				mechanisms surround their use.			
	Future of AI Act[345] (Introduced in 2017, projected to become law by the end of 2018)	**US**	Bill	**The bill was introduced in order to set up an interdisciplinary advisory committee to provide insights and recommendations on an appropriate regulatory framework for AI.**	**The law shall contribute to more informed decisions being taken on AI and algorithmic decision-making through an interdisciplinary advisory committee.**[346]	**Issued by the federal government.**	**It affects the industry, academia, NGO.**
	Senate Bill No. 1001[347] (Introduced in 2018, projected to become law in 2019)	**California**	Bill	**The law stipulates that "AI bots" shall not communicate with individuals in California online with the intent to mislead the other person about its artificial identity for the purpose of knowingly deceiving the person about the content of the communication in order to**	**The law aims to prevent that bots are able to influence decisions of individuals.**[348]	**Issued by the Californian Government.**	**Targeted at civil society, industry and the general population.**

[345] See https://www.congress.gov/bill/115th-congress/house-bill/4625/text
[346] Ibid.
[347] Senate Bill No. 1001, Chapter 892, 17th Congress (2018), p.1, Available at https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001
[348] Ibid.

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				incentivise a purchase or sale of goods or services in a commercial transaction or to influence a vote in an election.			
Singapore	**Advisory Council on Ethical Use of AI and Data**[349] (Initiated in 2018)	**Singapore**	Institution	**The Advisory Council will assist the Government to develop ethics standards and reference governance frameworks and publish advisory guidelines, practical guidance, and/or codes of practice for the voluntary adoption by the industry.**	**The Council aims to provide advice on how the private sector shall deal with the uptake of AI.**	**Issued by the government.**	**Targeted at the private sector and industry**
	Discussion paper on responsible development and adoption of AI[350] (Issued in 2018)	**Singapore**	Discussion paper	**Singapore's Personal Data Protection Commission (PDPC) has issued the paper which is based on two principles: Decisions made by or with the assistance of AI**	**The aim of the paper is to establish a baseline for discussion across stakeholders, encouraging common definitions.**	**Government, industry, civil society and academia involved in the composition of the discussion paper.**	**Targeted at consumers and users of AI systems**

[349] See https://euagenda.eu/upload/publications/untitled-128126-ea.pdf
[350] See https://www.pdpc.gov.sg/Resources/Discussion-Paper-on-AI-and-Personal-Data

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				should be explainable, transparent and fair to consumers. AI systems, robots and decisions should be human-centric.			
United Nations	**United Nations Activities on Artificial Intelligence**[351] (Issued in 2018)	**Worldwide**	Report	**The document contains separate reports of many UN agencies regarding their activities in the field of Artificial Intelligence. Many agencies have established departments which are working specifically on potential risks and benefits of AI in a variety of fields, ranging from labour to health.**	**Worldwide transformations of the economy and labour due to the application of AI. Development of predictive tools for health and natural threats. Forecasting of crime activity etc.**[352]	**The report was created by the UN agencies.**	**It affects many aspects of society worldwide in the fields of education, labour, health etc.**
	AI for Good series (Initiated in 2017)	**Worldwide**	Summit	**identify practical applications of AI and supporting strategies to improve the quality and**	**Identify the potential impacts of AI and machine learning on society.**[353]	**Organised by the International Telecommunications Union (ITU), the Association for Computing**	**A variety of members of society are affected by topics discussed at the event.**

[351] United Nations Activities on Artificial Intelligence (2018) available at: https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2018-1-PDF-E.pdf
[352] Ibid.
[353] See https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				sustainability of life on our planet covering issues such as the explainability, transparency and robustness of AI algorithms.		**Machinery (ACM) and various UN sister agencies. Speakers included many experts from academia, industry, government and civil society.**	
	Centre on Artificial Intelligence and Robotics (Launched in 2017 in The Hague)[354]	**Worldwide**	Research Centre	**Focussed on the impacts of using AI in areas such as crime prevention, criminal justice, law enforcement and national security, as well as the legal, social and ethical concerns involved.**	**Understanding and addressing the risks and benefits of AI and robotics from the perspective of crime and security through awareness-raising, education, exchange of information, and harmonization of stakeholders.[355]**	**Network of stakeholders including the International Criminal Police Organization (INTERPOL), the ITU, the Institute of Electrical and Electronics Engineers (IEEE), the Foundation for Responsible Robotics, the World Economic Forum, Centre for Future Intelligence (CFI) and others**	**Public and private stakeholders involved in crime prevention, criminal justice, law enforcement and national security.**
Organisation for Economic Co-operation and Development (OECD)	**Artificial Intelligence expert group at the OECD (AIGO)**	**OECD member states**	Expert group	**Ensuring AI does not exacerbate inequality, as well as working towards the mitigation of**	**Assisting governments, business, labour and the public to maximise the benefits of AI,**	**Experts from OECD member countries and think tanks, business, civil society and**	**Aims at addressing a variety of issues of large parts of society and business.**

[354] See http://www.unicri.it/news/article/2017-09-07_Establishment_of_the_UNICRI
[355] See http://www.unicri.it/topics/ai_robotics/centre/

Country / Region	Initiative[275] (Timeline)	Coverage[276]	Type	Material Scope	Expected Impact[277]	Initiated by	Affected Stakeholders
				biases and ensuring the safety, security, transparency and accountability of AI.	**while keeping an eye on ethical progress of AI and minimising risks.[356]**	**labour associations and other international organisations.**	

[356] See http://www.oecd.org/going-digital/ai/oecd-creates-expert-group-to-foster-trust-in-artificial-intelligence.htm

5.1 EU level

The use and applicability of algorithms are not governed by an overarching legislative framework at the EU level. Instead, there are sector-specific legislative instruments that govern the issue, some explicitly, others implicitly. Furthermore, in recent years, several policy documents have discussed the issue of algorithmic decision-making and funding mechanisms have provided funding for the research on algorithms.

The General Data Protection Regulation (GDPR)[357] makes a direct reference to algorithmic decision-making by stating that a "data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."[358] This prohibition can be lifted if an individual has given explicit consent.

Other EU legislative acts have been implemented to govern and regulate the activity of algorithms, such as the Markets in Financial Instruments Directive (MiFID II[359]), which focuses on the use of algorithms in the financial markets and regulates algorithmic trading. Specifically, the Directive imposes reporting and transparency requirements to companies engaging in algorithmic trading toward authorities in Member States, while also imposing an obligation of investment firms to have 'risk controls and effective systems' in place to ensure the stability and resilience of the financial system[360]. In particular, the directive specifies that investment firms disclose the nature of their algorithmic trading strategy, which not only leads to a more transparent procedure, but also increases the **accountability** between clients, authorities and trading companies. The European Securities and Market Authority (ESMA), has the responsibility to inform the public about the project, produce reports, register the discussed trading venues and monitor the implementation of the provision.[361]

The European Parliament has acknowledged the need to define a regulatory landscape for the use of algorithms in order to assess the potential impacts of algorithms on society. As noted by the European Parliament, algorithmic transparency and end-user awareness are needed to uphold democracy, citizen trust, fair competition, and stimulate innovation in digital societies.[362]

In addition, in its 2016 Communication on *Online Platforms and the Digital Single Market – Opportunities and Challenges for Europe*[363], the Commission has recognised that online platforms play an important role in the future of the EU economy, and that a balanced regulatory framework for online platforms in the digital single market is a necessity. This call

---

[357] 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)
[358] Article 22, GDPR
[359] Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments amending Directive 2002/92/EC and Directive 2011/61/EU
[360] Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments amending Directive 2002/92/EC and Directive 2011/61/EU
[361] See https://www.esma.europa.eu/policy-rules/mifid-ii-and-mifir
[362] http://www.europarl.europa.eu/sides/getDoc.do?type=COMPARL&reference=PE-587.719&format=PDF&language=EN&secondRef=01
[363] European Commission COM(2016) 288 final (2016), *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Online Platforms and the Digital Single Market Opportunities and Challenges for Europe*

for a balanced regulatory framework also considers the need to foster the innovation brought by online platforms, noting that any future regulatory measures proposed at EU level should only address clearly identified problems relating to a specific type or activity of these platforms in line with better regulation principles. Generally, the Commission notes specific key principles for the design of a balanced framework relating to the use and activity of online platforms at EU level, such as a level playing field for comparable digital services, and a responsible behaviour for online platforms to protect core values[364] Those were also discussed in a public consultation of the European Commission[365]. In addition, a string of European algorithmic defamation cases, particularly involving search results[366] have also compelled, in some jurisdictions, search engines to remove defamatory results from their listings.

More recently, the European Commission Communication "Artificial Intelligence for Europe" acknowledged that an alignment between all Member States through the drafting of a **coordinated plan on AI** is fundamental in order to leverage on the benefits of AI whilst mitigating its risks and in order to gain a competitive advantage on the international level[367]. The latter is particularly relevant when considering the recent developments in the AI global competitive landscape, including the AI strategic coordination and / or investment plans presented by other countries (i.e. including national-level approaches taken by Member States).[368,369] In its Communication the Commission highlighted that any approach to AI shall capitalise on the EU's key strengths in relation to the uptake of AI, including researchers, the digital single market and a wealth of industrial, research and public sector data.

Whilst ensuring that these aspects are leveraged upon, the Communication highlights that any approach to AI needs to **empower individuals and consumers**. It is important that when interacting with "an automated system, consideration should be given to when users should be informed on how to reach a human and how to ensure that a system's decisions can be checked or corrected." Respectively, the European Commission:

- "Has established the European AI Alliance to develop draft AI ethics guidelines, with due regard to fundamental rights, in cooperation with the European Group on Ethics in Science and New Technologies;
- Will publish, by mid-2019, a report on the broader implications for, potential gaps in and orientations for, the liability and safety frameworks for AI, Internet of Things and robotics;
- Will support national and EU-level consumer organisations and data protection supervising authorities in building an understanding of AI-powered applications with the

[364] For a full list of these key principles, see 'European Commission COM(2016) 288 final (2016), Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Online Platforms and the Digital Single Market Opportunities and Challenges for Europe'

[365] European Commission (2015 - 2016) Public consultation on the regulatory environment for platforms, online intermediaries, data and cloud computing and the collaborative economy

[366] Diakopoulos, N. (2013) The Case of the Shameless Autocomplete. Available at http://towcenter.org/algorithmic-defamation-the-case-of-the-shameless-autocomplete/

[367] European Commission (2018) 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, SWD(2018) 137 FINAL

[368] For an extended overview of the AI strategies outlined herein, see https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

[369] European Commission (2018) 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, SWD(2018) 137 FINAL

input of the European Consumer Consultative Group and of the European Data Protection Board."[370]

The Commission also made a regulatory proposal for promoting fairness and transparency for business users of online intermediation services.[371] It specifies that providers of online intermediation services should "outline the main parameters determining ranking beforehand, in order to improve predictability for business users", as it has a high impact on the commercial success of the businesses.[372] Finally, the proposed legislation should outline how businesses can actively influence ranking against remuneration and how the ranking mechanism takes the characteristics of the actual goods or services into account. Such measures offer a practical policy scenario to address the key issues of under transparency and scrutiny identified in the academic literature review above.

5.2 Selected EU Member States

Estonia

Estonia is one of the most digitalised countries in the world and has facilitated many bureaucratic procedures by the introduction of e-government.[373] Administration process automation can be seen as a strategic necessity for the implementation of those e-government procedures, which include voting, registration etc.[374]

Estonia's Personal Data Protection Act, which has been updated in 2008, specifies that a decision by a data processing system which results in legal consequences for the data subject or significantly affects him or her, is prohibited except for a few cases.[375] Namely, this includes a right to object against the decisions made by an automated process and the provision that "legitimate interests" of the data subject were protected. The article also emphasises that the affected subject must be informed in an understandable matter of the process. Similar provisions were later specified in the recently issued EU-wide GDPR.

Recently, a debate has been incited whether the liability issue that is arising with algorithm-based decision-making should be solved by giving algorithms a separate legal status, which would be similar to companies.[376] This so-called *Kratt law*, a name deriving from an Estonian folklore figure, should provide a legal framework on how to hold an AI or algorithm responsible for actions, in case of punishable behaviour or actions.[377] In March 2018, Estonia announced the launch of an expert group, which will comprise state authorities, universities, companies and independent experts, to investigate the question of responsibility for decisions made by systems.[378] The desired outcome of these expert consultations is projected to be a bill which allows the real-life application of fully autonomous information systems, a clear legal

[370] Ibid., p. 16.
[371] European Commission (2018) Proposal for a Regulation of the European Parliament and of the Council on promoting fairness and transparency for business users of online intermediation services
[372] Ibid, p. 15.
[373] See https://e-estonia.com/
[374] Priisalu and Ottis (2017) "Personal control of privacy and data: Estonian experience" Health Technology 7, pp441-451 https://link.springer.com/content/pdf/10.1007%2Fs12553-017-0195-1.pdf
[375] See https://www.riigiteataja.ee/en/eli/512112013011/consolide
[376] See https://e-estonia.com/ai-and-the-kratt-momentum/
[377] See https://www.newyorker.com/magazine/2017/12/18/estonia-the-digital-republic
[378] See https://www.riigikantselei.ee/en/news/estonia-will-have-artificial-intelligence-strategy

framework regarding their decisions as well as a sufficient concept of supervision. As even the creators of algorithms often cannot fully explain why a decision has been made by an AI, the last point is in particular difficult to reach.[379]

Policy opinion by the authors

As already discussed in the academic debate section about liability of algorithmic systems and objects operated by computational systems based on deep learning, there is no uniform legal regime in place in Europe, although there have been calls by MEPs to implement a special legal status for AI and AI operated machines in the long run.[380][381] Instead, some countries (e.g. Germany) have a solution that absolved owners of an algorithmic system of liability if they have taken sufficient care, while others (e.g. France) apply regimes of strict liability, where the owner is always liable when damage is caused. The *Kratt law* proposes an alternative model which could allow for an operation of fully autonomous information systems with a clear framework regarding liability in combination with a sufficient concept of supervision.

France

France was one of the first countries in the world to regulate automated decision-making in the late 1970s, prohibiting solely automated significant decisions across a range of sectors.[382] While this law has been adapted and now forms part of the Europe-wide data protection framework, the country has extended initiatives around this area beyond that in many other Member States.

The Digital Republic Act 2016[383] contains particular provisions on 'algorithmic treatment' of individuals by the public administration. It places requirements on government bodies that make decisions solely or partially by algorithmic systems to give individuals certain transparency rights regarding which decisions were made. These provisions were extended by decree in March 2017 and require public institutions to make the following issues transparent:

- The degree and the mode of contribution of the algorithmic processing to the decision-making;
- The data processed and its source;
- The treatment parameters and, where appropriate, their weighting, applied to the situation of the person concerned; and
- The operations carried out by the treatment.[384]

In order to adapt the French law to the GDPR, the National Assembly adopted the New Personal Data Protection Act in May 2018. [385] The law specifies in detail to which extent algorithmic decision-making for legally binding decisions is permitted. Although the wording of Article 21

[379] Ibid.
[380] See http://www.europarl.europa.eu/news/en/press-room/20170210IPR61808/robots-and-artificial-intelligence-meps-call-for-eu-wide-liability-rules
[381] See subsection Liability in the Academic Debate.
[382] Loi n° 78-17 du 7 janvier 1978 relative à l'informatique, aux fichers et aux libertés, articles 2–3.
[383] Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.
[384] Edwards, Lilian and Michael Veale (2018) "Enslaving the Algorithm: From a "right to an explanation" to a "right to better decisions"?" IEEE Security & Privacy 16(3) pp46–54 Available at https://doi.org/10.1109/MSP.2018.2701152
[385] See https://blogdroiteuropeen.files.wordpress.com/2018/06/olivia.pdf

of the national Personal Data Protection Act prohibits decisions based solely on automated processes, there are some exceptions for a systematic use for administrative individual decisions. A prerequisite for such processes is the explicit reference of its automated-processing decision nature and the possibility of gaining ex-post explanation of the processes, which includes the possibility to look into the source code. Applying these processes to sensitive data remains prohibited. In the academic debate, it is argued that Art. 21 of the New Data Protection Act, adapting the necessity of "suitable measures to safeguard the data subject's right and freedoms and legitimate interests" of Art. 22 (2) b) of the GDPR, also implies the right of a human intervention in an automated process.[386] One of the first applications of the new provisions regarding automated decision-making was the introduction of the *Parcoursup* platform, which has been introduced in 2018 in order to process the applications of prospective undergraduate students to public universities.[387] In its first year, data of some 900,000 aspiring students have been processed through automated decision-making.[388] Different to the formerly applied platform *APB*, *Parcoursup* offered the students the necessary possibility of human intervention.

Policy example by the authors

To give an example for a real-life application, a candidate who has received a rejection by the algorithm applied by Parcoursup can receive information regarding the process that led to the negative decision. According to Art. L-612-3 of the code of Education, the law provides a sufficient level of information if "the candidates are informed of the possibility to obtain the communication of the information regarding the criterion and the modalities of their applications and the pedagogical grounds of the final decision".[389] Like pointed out in the academic debate section discussing the aspects of **transparency** and **scrutiny** of algorithm-based decision-making, these provisions might sound attractive and have significant public support, but concerns have been raised by some authors that they might only "provide a meaningless, non-actionable form of explanation that does little more to help deal with algorithmic harms than privacy policies individuals have little time to read."[390] Especially looking into a very exhaustive and non-comprehensive source code might not be a practicable option on the individual level.

Beyond the Digital Republic Act, in 2018, the government presented its €1.5 billion plan[391] to transform France into a global leader in AI research, training, and industry at the *AI For Humanity Summit*[392]. The plan is extensively based on Cédric Villani's report "*For a Meaningful Artificial Intelligence: Towards a French and European Strategy*"[393] and consists of the following main components:

- Development of specific initiatives to strengthen France's AI ecosystem and attract international level investment;

[386] Ibid, p.6
[387] See https://www.parcoursup.fr/index.php?desc=essentiel
[388] See https://blogdroiteuropeen.files.wordpress.com/2018/06/olivia.pdf
[389] Ibid, p.8
[390] See the chapter on Transparency and Scrutiny in the section Academic Debate.
[391] République Française – Le Gouvernement (2018) *Rapport de synthèse – France Intelligence Artificielle. See for example:* Available at https://www.gouvernement.fr/en/artificial-intelligence-making-france-a-leade
[392] See https://www.aiforhumanity.fr/en/
[393] See https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

- Implementation of an open data policy aiming to drive the application and adoption of AI in strategic sectors for France, particularly healthcare, transportation, environment, and defence;
- Creation of a financial and regulatory framework to support the development of 'AI Champions';
- Development of an ethical framework for AI, in order to ensure that its development and use is transparent, explainable, and non-discriminatory;

In addition to the political strategy as well as legislative initiatives, also institutional steps have been taken to increase the transparency of algorithmic decision-making. Organisations such as the French Digital Council, founded in 2011, already provide an advisory service to government.[394] [395]The Council is responsible for studying digital issues, in particular the issues and prospects for the digital transition of society, the economy, organizations, public action and territories. The Council is composed of thirty voluntary experts representing the digital economy, academia and civil society.

Germany

Germany's AI strategy will be published in December 2018 at the Digital Summit in Nuremberg. However, the German federal cabinet has already released a white paper outlining the main goals of this strategy[396], which are the strengthening and expansion of German research in AI, the establishment of research and development collaboration, and the creation of measures to attract international talent. In addition to those primary target areas, the strategy also included considerations on **the importance of algorithmic transparency**, auditability and control in decision-making to ensure that no discrimination or manipulation is possible when algorithms take decisions or model forecasts. In the following months, the government plans to consult stakeholders across Germany to further develop the strategy on those points.

In addition, the German government has announced a new commission, the **Enquete Commission on AI**, which aims to investigate the societal impact of AI and algorithmic decision-making. The Enquete Commission is going to investigate opportunities and challenges of algorithmic decision-making and will formulate recommendations for action which will be presented in autumn 2020. The Commission consists of 19 MPs and 19 AI experts[397]. The first sessions of the commission revealed that partisan differences exist regarding potential advantages and disadvantages of AI in everyday life and the economy.[398] While the right-wing AfD (Alternative für Deutschland) emphasised the potential for state surveillance, parties like the conservative CDU (Christlich-Demokratische Union) and the liberal FDP (Freie Demokratische Partei) underpin the potential economic benefits (while underlining the importance of preserving individual **privacy**). Finally, the social-democratic party SPD

[394] See French presidential decree n°2011-476.
[395] *Conseil National du Numérique* (CNNum), whose mission and objectives are listed at https://cnnumerique.fr/missions/
[396] See https://www.bmwi.de/Redaktion/DE/Downloads/E/eckpunktepapier-ki.pdf?__blob=publicationFile&v=4
[397] See https://www.bundestag.de/dokumente/textarchiv/2018/kw26-de-enquete-kommission-kuenstliche-intelligenz/560330
[398] See https://www.bundestag.de/dokumente/textarchiv/2018/kw39-pa-enquete-kuenstliche-intelligenz/567956

(Sozialdemokratische Partei Deutschlands) underpins the importance of sharing the economic benefits of algorithm-based machines among the social hemisphere, while the Greens (Bündnis 90/Die Grünen) and the leftist Die Linke warn of potential social and regulatory problems arising with AI.[399]

Last but not least, Germany has also implemented a series of policy measures with the aim of further developing AI in the country more generally. These measures have the overarching objective of ensuring a collaborative partnership between stakeholders in government, academia, and industry in the context of integrating AI technologies and applications into Germany's key export sectors[400]. To this end, the *German Centre for AI (DFKI)*[401], the *Alexander von Humboldt Foundation*[402], and the *Plattform Lernende Systeme*[403] play critical roles.

Italy

Italy's AI strategy was published in the form of a white paper (i.e. *Artificial Intelligence: At the Service of Citizens*[404]) in March 2018. The strategy focuses on outlining the challenges of AI as well as providing recommendations on how to overcome those challenges. As many other national strategic papers, the white paper focuses on how to encourage the research and uptake of AI. More specifically, the paper recommends measures on education and training, the creation of a national platform to promote the collection of annotated data, as well as the creation of a National Competence Centre and a Trans-disciplinary Centre on AI. In addition, the paper outlines the application of AI in the public administration.

The paper also **highlights existing risks of algorithmic decision-making such as** the disempowering of people, erroneous decisions, unfair or discriminatory consequences of those decisions, the violation of privacy of those subject to algorithmic decision-making, and the risk of unwanted conditioning of those subject to algorithmic decision-making.[405] The paper thus highlights that the use of algorithms in decision-making processes related to social, health and judicial issues therefore requires a thorough ethical reflection and more broadly a governance reflection. The strategy highlights that key aspects of this governance debate are the following:

- Data quality and neutrality;
- Responsibility through accountability/liability;
- Transparency and openness;
- Protection of the private sphere.

To address the challenges in relation to each of these aspects, an anthropocentric approach is suggested where the use of AI and algorithms must always be at the service of people and not vice versa. The paper emphasises the need of a certain catalogue of criteria for public services operated on AI and states that the "criteria to be used undoubtedly include transparency of

[399] Ibid.
[400] See: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/oral/75597.html
[401] See https://www.dfki.de/web
[402] See https://www.humboldt-foundation.de/web/home.html
[403] See https://www.plattform-lernende-systeme.de/home-en.html
[404] See https://ai-white-paper.readthedocs.io/en/latest/, see also: https://ia.italia.it/en/task-force/
[405] Ibid., Challenge 1

the algorithms, the construction logic of the databases on which they operate and defining the related responsibilities". [406]

In September 2017, the Agency for Digital Italy (AGID) has established an AI Task Force.[407] The task force includes 30 direct members and around 450 community members from many sectors. The aims of the Task Force include:

- The analysis of the use of AI in public administration,
- Defining the potential of AI services and applications, and
- Exploring the ethical and social implications of the use of AI.

As part of its mandate, the AI Task Force has initiated the Observatory on Artificial Intelligence, which aims to analyse AI-related public conversations on social networks through technology.

Poland

The development of an AI strategy is still under development in Poland, as its government held a first round of discussions in this regard in May 2018[408]. This roundtable discussion focused on the tools and policies needed to foster an environment that will enable the creation and development of AI technologies in Poland.

During the roundtable it was stressed that a Polish AI strategy will include considerations on health care, public administration, education and cyber security. According to Deputy Prime Minister Jarosław Gowin the government would start working on an AI strategy in autumn 2018 by consulting a group of technological leaders and scientists.[409]

Whilst not having developed an own AI Strategy, Poland has been considered the initiator of the Visegrád's Group[410] position on AI.[411] In its position on AI, the Visegrád group mentioned that among others the use of AI in decision-making in the public administration shall be prioritised as well as issues on cybersecurity and trust more generally. While the position paper highlights also legal, social and ethical considerations these are not tailored to algorithmic decision-making only.[412]

Spain

In November 2017, Spain has assembled a commission of experts and "sages" who are currently working on an ethics code for the societal, judicial and ethical implications that come with the use of artificial intelligence and Big Data in the private and public sector as well as in

[406] See https://ai-white-paper.readthedocs.io/en/latest/doc/capitolo_3_sfida_5.html
[407] See https://futureoflife.org/ai-policy-italy/
[408] See https://www.money.pl/gospodarka/wiadomosci/artykul/rewolucyjny-plan-dla-polski-powstaje,205,0,2406605.html
[409] http://scienceinpoland.pap.pl/en/news/news%2C30610%2Cexpert-implementing-ai-development-strategy-must.html
[410] The Visegrád Group (Czechia, Hungary, Poland and Slovakia) is the cooperation of the Central European region in a number of fields of common interest within the all-European integration.
[411] https://news.microsoft.com/uploads/prod/sites/58/2018/06/Artificial_Intelligence.pdf
[412] http://digiczech.eu/wp-content/uploads/2018/04/V4_NON_PAPER_ON_AI_09_04_2018.pdf

society in general.[413] The group consists of leading figures in industries, academia and the non-governmental sector.

The conclusions of this commission, manifested as a white book, includes a report over the growing use of data in public administration and in companies. The recommendations include an ethics code for the use of data in public administration, as well as a code of conduct for companies regarding the use of Artificial intelligence and data.

This initiative is part of the Digital strategy for "España Inteligente 2025" that the Spanish Government has issued in 2017.[414] The first pillar of this strategy includes the economy and society of data, where conclusions and proposals of the white book and the draft of the codes of conduct for public administration as well as the private sector are elaborated.

United Kingdom

The United Kingdom has had several recent initiatives aimed at considering governance of automated decision-making and artificial intelligence.

Several new bodies have been created within and around government. A report from the Royal Society and British Academy recommended the creation of a 'data stewardship body'[415] designed to ensure that certain high-level principles around data use were met in the entire data ecosystem:

- Data use should promote human flourishing;
- Protect individual and collective rights and interests;
- Ensure that trade-offs affected by data management and data use are made transparently, accountably and inclusively;
- Seek out good practices and learn from success and failure;
- Enhance existing democratic governance.

It would do so through i) anticipation, monitoring and evaluation; ii) building practices and setting standards; and iii) clarifying and enforcing rules and remedying harms.

This data stewardship recommendation in part led the UK Government to establish a body called the *Centre for Data Ethics and Innovation*, an arms-length public body attached to the Department for Digital, Culture, Media and Sport (DCMS). The role this advisory body will play will be shaped in the coming years.[416] However, in initial communications,[417] the Centre has been tasked with:

i. Analysing, as well as anticipating gaps in governance;
ii. Detailing best practices related to the ethical and innovative uses of data; and

---

[413] https://www.mincotur.gob.es/es-ES/GabinetePrensa/NotasPrensa/2017/Paginas/grupo-expertos-big-data20171114.aspx

[414] https://www.lavanguardia.com/politica/20170628/423744247082/el-gobierno-impulsara-una-estrategia-digital-para-espana-en-el-horizonte-2025.html

[415] The Royal Society and the British Academy (2017) *Data Management and Use: Governance in the 21st Century*. London.

[416] Department for Digital, Culture, Media and Sport (2018) Consultation on the Centre for Data Ethics and Innovation. HM Government, London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/715760/CDEI_consultation__1_.pdf

[417] Ibid.

iii. Advising the UK Government on policy or regulatory needs.

More specifically, the public consultation launched prior to the Centre establishes a framework of six high-level themes related to the ethical and innovative use of data-driven and AI technologies, strongly reflecting the themes discussed throughout this report. These themes are:

- **Targeting** and its potential to restrict the information and choices available to users, as well as influence, manipulate or control behaviour.
- **Fairness** and the potential for bias and discrimination to be present in AI systems.
- **Transparency** and the need for interpretability and explainability.
- **Liability** and its role in distributing risks and benefits.
- **Data access** and the need to establish sound incentives and structures for the creation, collection, analysis, sharing and trading of data.
- **Intellectual Property (IP) and ownership**, highlighting the need for IP regimes to keep up with new innovations and for questions related to the ownership of data and innovation to be answered.

In addition to this, there has been the creation of several other units attached to government, including the Government Office for AI, a business-led AI Council, as well as a forthcoming National Data Strategy.[418] There has also been additional investment in the Alan Turing Institute, a company established with seed funding from government to connect university expertise in the UK, and in the Information Commissioner's Office (the national data protection regulator), who have released a technology strategy for engagement around AI.[419] Several parliamentary inquiries have focussed on the issue of algorithms, including the House of Commons Science and Technology Committee's three inquiries 'Algorithms in Decision-Making'[420], 'The Big Data Dilemma'[421] and 'Robotics and Artificial Intelligence'[422] and the work of the House of Lords Select Committee on Artificial Intelligence.[423]

Netherlands

According to Bendert Zevenbergen, Research Fellow at the Center for Information Technology Policy at Princeton University, "the Netherlands is carefully and slowly building its foundations

[418] Press release Department for Digital, Culture, Media & Sport (2018) World-leading expert Demis Hassabis to advise new Government Office for Artificial Intelligence. HM Government, London. https://www.gov.uk/government/news/world-leading-expert-demis-hassabis-to-advise-new-government-office-for-artificial-intelligence

[419] Information Commissioner's Office (2018) Technology Strategy 2018-2021, Wilmslow. https://ico.org.uk/media/about-the-ico/documents/2258299/ico-technology-strategy-2018-2021.pdf

[420] House of Commons Science and Technology Committee (2018) https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2015/inquiry9/

[421] House of Commons Science and Technology Committee (2016) https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2015/big-data/

[422] House of Commons Science and Technology Committee (2016) https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2015/robotics-and-artificial-intelligence-inquiry-15-16/

[423] House of Lords Select Committee on Artificial Intelligence (2018) https://www.parliament.uk/ai-committee

for an AI policy and strategy".[424] However, the country is still in the stage of intensive research rather than policy implementation. In contrast to countries like France, Germany and the UK, which invest heavily in AI research, the Netherlands faces difficulties to keep up, especially as there exists a scarcity of AI researchers. The government's focus at the current stage seems to be on understanding potential regulatory approaches to AI and algorithmic decision-making through research on issues such as algorithmic decision-making and social values and human rights. In addition, some collaborative projects are created such as a project of the University of Amsterdam together with the municipality and other businesses to create Amsterdam's AI Hub.[425] Organisations active in AI governance debates in the Netherlands include the Rathenau Instituut and the Scientific Council for Government Policy (WRR).

Ireland

Ireland has a vivid scene of start-up and established companies focusing on AI. Therefore, many initiatives have focused on encouraging research both among industry stakeholder and encouraging collaboration between industry and academia.[426] The debate on AI and algorithmic decision-making is rather driven by business and academia, than by the Government. A national strategy, which has been initiated or developed by other countries like France, Britain and South Korea already, has not yet been adopted by the Irish Government.[427]

Despite criticism by the Irish business association (ibec), which urged to establish more ambitious policies and guidelines in the field of AI and algorithmic decision-making, a national strategy on those topics has not yet been initiated by the government. [428][429]

As a part of the AI strategy developed by the Irish Economic Development Agency (IDA) and Enterprise Ireland, the country has recently launched a national Masters in Artificial Intelligence, which is entirely industry-driven.[430] It is also worth noting that the Irish Data Protection Commission is the lead authority for many global technology firms based in the country, and therefore regulatory initiatives and decisions in Ireland are likely to be important in the context of the governance of algorithmic systems and platforms.

Denmark

Despite not being specifically focused on algorithmic decision-making and its consequences, Denmark's *Strategy for Digital Growth*[431] focuses on creating growth and wealth for Danish citizens through advances in the Internet of Things, big data and AI. Apart from that the Danish

---

[424] See https://www.considerati.com/publications/ai-policy-for-the-netherlands.html
[425] See: http://amsterdamdatascience.nl/news/uva-the-city-of-amsterdam-commit-to-artificial-intelligence-hub-at-science-park/
[426] See e.g. https://www.irishtimes.com/business/innovation/ireland-needs-a-national-strategy-for-artificial-intelligence-1.3545739
[427] Ibid.
[428] Ibid.
[429] See https://www.ibec.ie/IBEC/Press/PressPublicationsdoclib3.nsf/vPages/Newsroom~calls-for-government-to-make-ireland-global-digital-leader-26-04-2018?OpenDocument
[430] European Commission (2018) Report on The European Artificial Intelligence landscape
[431] Ministry of Industry, Business and Financial Affairs (2018) Press release: New Strategy to make Denmark the New Digital Frontrunner, https://eng.em.dk/news/2018/januar/new-strategy-to-make-denmark-the-new-digital-frontrunner/

Government has established an expert group on data ethics which will develop a set of recommendations on how to treat data.[432] The Danish government has also recently announced renewed funding for research in AI for 2019.[433]

Finland

Finland has appointed a steering group in 2017 to examine how the country could become a leader in the application of AI technologies. The final work from this steering group will only be published in 2019. However, it has since published two reports. The first report was mainly targeted on economic growth through AI, which is discussed in eight proposals for an ethical and beneficial integration of AI into the Finnish society. Furthermore, potential strategies to make Finland a forerunner of research regarding AI are elaborated.[434] The second report "Work in the Age of Artificial Intelligence" introduces further policy recommendations in the context of the ethical implementation of AI processes into the workplace and a voucher system for lifelong learning, which should help employees to adapt to changes due to the implementation of AI processes in their workplaces. [435]

Nordic Council of Ministers

The Nordic Council of Ministers for Digitalisation[436] announced in 2017 that it wants the Nordic and Baltic region to be a digital leader and show the way in Europe. At the same time, the Council is eager to see core Nordic values (freedom of speech, openness, democracy, shared values and equal rights) to be respected when AI is gaining a bigger role in society. Therefore, the ministers and the Nordic Council of Ministers for Digitalisation are drawing up ethical guidelines, standards, principles and values for when and how AI should be deployed.[437]

5.3 Third countries

Australia

The *Australian Competition & Consumer Commission* (ACCC) has released a call for opinions of consumers, media organisations, digital platforms, advertising agencies and advertisers, which are supposed to be included in an inquiry about the power of near monopolist digital platforms.[438] The ACC is seeking feedback on the impact of the potential bargaining power of

---

[432] See https://em.dk/english/news/2018/03-12-new-expert-group-on-data-ethics
[433] See https://www.bt.dk/politik/regeringen-vil-satse-paa-forskning-i-kunstig-intelligens
[434] Ministry of Economic Affairs and Employment (2017) Publication: Finland's Age of Artificial Intelligence, FI Government, Helsinki. http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf?sequence=1&isAllowed=y
[435] Ministry of Economic Affairs and Employment (2018) Publication: Work in the age of artificial intelligence, FI Government, Helsinki. http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160980/TEMjul_21_2018_Work_in_the_age.pdf
[436] https://www.norden.org/en/information/about-nordic-council-ministers-digitalisation-mr-digital
[437] See https://www.norden.org/en/information/nordic-trust-meets-artificial-intelligence
[438] Australian Competition & Consumer Commission (2018) Media Release: ACCC Seeking views on news and digital platforms inquiry. https://www.accc.gov.au/media-release/accc-seeking-views-on-news-and-digital-platforms-inquiry

these digital platforms (like Google for search engines and Facebook for Social Networks) on the quality and the emphasis of news content of broadcasters, which are dependent on visibility on these platforms. In its preliminary report, which is supposed to be released in December 2018, the ACCC also want to discuss how the use of algorithms affects the presentation of news for digital platform users.[439] The Issues Paper released in February 2018 also discusses to what degree digital platforms are transparent about the applied algorithms and whether they publicly inform content creators about significant changes within those algorithms.[440] This is considered the first major open-ended public inquiry of a major competition regulator and could lead to recommendations regarding legislative change, but primarily might lead to a better understanding of digital platforms in order to optimise future competition policy.[441] In particular, the findings of the future report might lead to a better understanding of **accountability**, which kind of behaviour on near monopolist digital platforms leads to what kind of outputs in KPIs for companies operating online (e.g. in traffic, visibility, leads to the website etc.)

Canada

As one of the first countries worldwide, Canada has introduced a national AI strategy called *Pan-Canadian Artificial Intelligence Strategy* in 2017.[442] The $125 million project has four major goals:[443]

- Increasing the number of outstanding artificial intelligence researchers and skilled graduates in the country.
- Establishing interconnected nodes of scientific excellence in Canada's main AI centres Edmonton, Montreal and Toronto.
- Developing global thought leadership on the economic, ethical, policy and legal implications of advances in artificial intelligence.
- Supporting a national research community on artificial intelligence.

Specific initiatives include establishing several AI institutes in the aforementioned AI prone cities, an AI & Society Program as well as a National AI Program.[444] The leading company of the project, CIFAR, expects the achievement of an enhancement of Canada's International profile in AI research and training, increased productivity in Canadian AI academic research and a better collaboration across geographic areas of excellence. Furthermore, the project aims at attracting and retaining Canadian AI talents within the country and making AI advances beneficial for a broader spectrum of the society in Canada.[445] The Government of Canada has also released draft guidance on Algorithmic Impact Assessments, and the Canada Treasury

[439] Ibid.
[440] Australian Competition & Consumer Commission (2018) Issues paper: Digital platforms inquiry. See https://www.accc.gov.au/focus-areas/inquiries/digital-platforms-inquiry/issues-paper
[441] See http://www.nortonrosefulbright.com/knowledge/publications/163857/world-first-inquiry-into-digital-platforms-in-the-media-sector
[442] See https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy
[443] Ibid.
[444] Ibid.
[445] Ibid.

Board has released a Directive on Automated Decision-Making under the authority of the Financial Administration Act. [446]

China

Through the implementation of its *New Generation AI Development Plan*[447] China aims to become the leading AI power by 2030 involving as many as 15 government agencies. The plan includes amongst other elements the use of AI to build a "safe and convenient intelligent society"[448]. In practice this includes the development of intelligent services (including: intelligent education, intelligent medical care, intelligent health and aged care), intelligent social governance (including: intelligent business, intelligent courts, intelligent city, intelligent transport, intelligent environmental protection) and the use of AI to protect public security. Furthermore, the plan also highlights that AI shall contribute to mutual trust and sharing in society.

In the context of the ambition to roll out AI across all societal sectors, the strategy also **acknowledges the need to establish laws, regulation and ethical norms surrounding the use of AI**. The plan highlights the need to carry out research on legal issues of AI in particular in the areas of civil and criminal liability, privacy, intellectual property. It is thus necessary to establish traceability and accountability systems and identify AI legal entity and related rights, obligations and responsibilities. The strategy further highlights that research will be carried out on setting up a framework on ethics and morals as well as a behaviour code for AI product research designers in order to regulate the potential AI risks such as robot alienation and safety. Furthermore, it will be necessary to exchange views on regulation and ethical guidelines internationally.[449]

Apart from the New Generation Plan, Progress in AI is strongly encouraged under the 13th Five-Year Plan released last year, as well as in notable state-driven industrial plans such as "Made in China 2025" which actively promote and support the development of advanced industries and technologies.[450]

India

India's national AI strategy[451], developed by the think tank NITI Ayagog[452], is not only directed at economic growth but also social inclusion. The strategy has three main objectives:

---

[446] https://canada-ca.github.io/digital-playbook-guide-numerique/views-vues/automated-decision-automatise/en/algorithmic-impact-assessment.html

[447] See http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

[448] China Science & Technology (2017) Newsletter: Next Generation Artificial Intelligence Development Plan Issued by the State Council. Department of International Cooperation Ministry of Science and Technology, P.R. China, Beijing, http://www.chinaembassy-fi.org/eng/kxjs/P020171025789108009001.pdf

[449] China Science & Technology (2017) Newsletter: Next Generation Artificial Intelligence Development Plan Issued by the State Council. Department of International Cooperation Ministry of Science and Technology, P.R. China, Beijing, http://www.chinaembassy-fi.org/eng/kxjs/P020171025789108009001.pdf

[450] See https://euagenda.eu/upload/publications/untitled-128126-ea.pdf

[451] See http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf

[452] See http://www.niti.gov.in/

- Enhance and empower citizens with the right skills to find high-quality employment opportunities;
- Invest in research and strategic sectors that can maximise economic growth and social impact and inclusion;
- Promote and support the scaling-up of Indian-made AI solutions to the rest of the developing world.

In this context, NITI Ayagog has provided the Indian government with over 30 policy recommendations to be implemented specifically in the sectors of healthcare, agriculture, education, smart cities, and smart mobility. The strategy acknowledges that automated decision-making is -similarly to human decision-making- not entirely infallible. The strategy mentions:

> *"one needs to be conscious of the potential vulnerabilities of our extant regulatory and societal structures which are dependent on human judgment and control, and thus subject to inherent biases and discrimination. Thus, to say that extant decision-making systems – individual, societal, regulatory or even judicial – are entirely devoid of these shortcomings would be a fallacy as these are dependent upon human limitations of knowledge, precedent, rationale and bias (explicit or subconscious)."*

The AI strategy mentions that **a problem in relation to fairness is a potentially biased dataset which informs the algorithm**. A solution suggested by the strategy is to identify the in-built biases and assess their (potential) impact, and in turn find ways to reduce the bias. While noting that this is a reactive approach, over time techniques may be identified that will bring neutrality to data feeding AI solutions, or build AI solutions that ensure neutrality despite inherent biases.[453]

Subsequently, the strategy also highlights the problem of **algorithmic transparency or the so-called "Black-Box Phenomenon"** which refers to the situation where very little is known on how algorithms take decisions. According to the strategy, the aim shall not be to "open the black box" since this is not in the interest of companies and is often not understood by the wider public. Instead "explainability" shall be the objective where individuals need to be able to understand how/why certain decisions are taken. Further collaborative research is needed on the ultimate form that "explainability" shall take.[454]

Based on the concerns on fairness and transparency, the AI strategy recommends the government to set up a consortium of Ethics Councils, Centres of Research Excellence (CORE)[455] and International Centers of Transformational AI (ICTAI)[456] to define the standard practices and monitor their adoption as well as the close cooperation with industry in different sectors.

---

[453] See http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf, p. 85.
[454] P. 86.
[455] The strategy suggests the setting up of "Centres of Research Excellence (CORE) focused on developing better understanding of existing core research and pushing technology frontiers through creation of new knowledge" (p.7)
[456] The strategy suggests "International Centers of Transformational AI (ICTAI) with a mandate of developing and deploying application-based research. Private sector collaboration is envisioned to be a key aspect of ICTAIs"

Japan

Japan was the second country to develop a national AI strategy. During the "Public-Private Dialogue towards Investment for the Future"[457] in April 2016, **the Strategic Council for AI Technology was established to develop a roadmap**. In March 2017, the Council announced the Artificial Intelligence Technology Strategy.[458] While discussing all main aspects in relation to AI, the strategy highlights that ethical aspects of AI technology (including: intellectual property rights, personal information protection, and promotion of open data) will be examined by the government as cross-sectional items.[459]

To supplement the Strategic Council's AI roadmap, an advisory expert group has been set up which is chaired by Dr. Osamu Sudoh and composed of experts from industry, academia and the private sector. The expert group **has drafted AI guidelines on the assessment of risks of AI** in each sector of society. The aim of the publication is to generate international discussions on the topic in forums such as the G7 or the OECD.[460]

The key message of the guidelines is that the aim should be to achieve a human-centred society where a balance is achieved between the benefits and risks of AI networks.[461] Furthermore, it shall be ensured that regulation/guidelines are technological neutrality and reviewed regularly. A few of the principles presented by the guidelines are relevant for algorithmic decision-making[462]:

- The principle of ethics (developers shall take necessary measures so as not to cause unfair discrimination resulting from prejudice included in the learning data of the AI systems);
- The principle of transparency (developers shall pay attention to the verifiability of inputs/outputs of AI systems and the explainability of their judgments);
- Principle of controllability (developers shall conduct verification and validation of AI systems in advance)
- Principle of privacy (developers shall make an effort to evaluate the risks of privacy infringements in advance)

South Korea

South Korea announced in March 2016 that it would invest EUR 760 million in AI technologies until 2021 mostly through the founding of a new public-private research centre bringing together large-scale technology conglomerates[463].

While not having produced an ethical code for AI in general or algorithmic decision-making processes, South Korea has developed an ethical code for robots. The government has adopted the "Intelligent robots development and distribution promotion act" which was last amended in 2016. The law defines "smart robot" as a mechanical device which perceives its external

[457] Prime Minister of Japan and His Cabinet (2016) New release: Public-Private Dialogue towards Investment for the Future, Tokyo, https://japan.kantei.go.jp/97_abe/actions/201604/12article6.html
[458] See http://www.nedo.go.jp/content/100865202.pdf
[459] P. 4
[460] See http://events.science-japon.org/dlai17/doc/MIC%20-%20France-Japan%20Symposium%2020171025.pdf
[461] See http://www.soumu.go.jp/main_content/000507517.pdf
[462] Ibid.
[463] European Commission (2018) Digital Transformation Monitor – USA-China-EU plans for AI: where do we stand?

environment, evaluates situations and moves by itself (Article 2(1))9. The act lays down requirements of certifications and regular quality control but does not specifically refer to decision-making processes of robots.

USA

The US was the first country to have started to implement a comprehensive AI research and development plan in May 2016 with the report 'Preparing for the Future of Artificial Intelligence,' and a companion 'National Artificial Intelligence Research and Development Strategic Plan'. The plan sets out a series of objectives for AI research for academia, industry and within government[464]. In relation to ethical concerns the strategy mentions that research needs to involve both understanding the ethical, legal, and social implications of AI, as well as developing methods for AI design that align with ethical, legal, and social principles. In relation to algorithmic decision-making two aspects are highlighted. First, the proper collection and use of data for AI systems is an important challenge. Second, it is an important question on how to achieve that the design of AI is inherently just, fair, transparent, and accountable.[465]

In addition to the federal initiatives, the New York City Council passed in December 2017 an **algorithmic accountability bill**. The law implements a task force that monitors the fairness and validity of algorithms used by municipal agencies.[466] The task force however does not have any special investigative powers or resources, and there is no obligation for policymakers to take up recommendations from the report it will provide.[467]

In addition to these strategic efforts and legislation, a bipartisan group of lawmakers established the *Artificial Intelligence Caucus* last year.[468] The purpose is to generate more dialogue between industry professionals and policy-makers as they collectively work to make technology safer, more sustainable and ethical. For example, recently the FUTURE of AI Act was introduced which aims to set up an interdisciplinary advisory committee to provide insights and recommendations on an appropriate regulatory framework for AI.[469] While the legislation is in its early stages, there is some optimism that it could become law by the end of 2018.

The Californian senate proposed a bill that makes it unlawful for AI bots to communicate with "another person in California online with the intent to mislead the other person about its artificial identity in order to incentivize a purchase or sale of goods and services in a commercial transaction or a vote in an election."[470] A bot is defined as an "automated online account where all or substantially all of the actions or posts of that account are not the result of a person" and the proposal is projected to become law in mid 2019.[471] The main incentive for the proposal by Senator Robert Hertzberg was the suspected interference of so-called "troll farms" in

[464] National Science and Technology Council (2016) The National Artificial Intelligence Research and Development Strategic Plan. Available at https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf
[465] See https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf p. 26.
[466] See https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0
[467] See https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0
[468] See https://artificialintelligencecaucus-delaney.house.gov/
[469] See https://www.congress.gov/bill/115th-congress/house-bill/4625/text
[470] Senate Bill No. 1001, Chapter 892, 17th Congress (2018), p.1, Available at https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001
[471] Ibid p. 1.

elections and controversial political discussions, like the #MeToo debate.[472] However, critics have raised concerns that the bill might infringe there first amendment right, which guarantees the freedom of expression.[473] [474]The United States' Department of Defense (DOD) also acknowledges the importance of Artificial Intelligence in Modern Warfare. The Pentagon's Defense Innovation Unit (DIU) has therefore made AI one of its five pillars of work.[475] Recently, the DOD has introduced the so-called *Joint Enterprise Defense Initiative (JEDI),* which aims at moving the departments computing system into a cloud.[476] Through this initiative, the Pentagon wants to "inject artificial intelligence into its data analysis and equip soldiers with real-time data during missions".[477] As Google has announced it will not bid for the project, as it might be against their AI principles, Amazon is considered the clear front-runner to close the deal. In addition to the JEDI initiative, the Pentagon has also introduced a project called *Maven*, which aims at applying machine learning to scan video footage in support of counterinsurgency and counterterrorism operations.[478] Currently, the Air Force is testing its "prototype warfare" strategy to analyse video surveillance footage in its AFRICOM mission in Africa.[479]

Singapore

In May 2017, a national initiative called AI.SG brought together several agencies to develop Singapore's deep AI capabilities. AI.SG aims to address with the help of AI major challenges affecting society and industry, invest in readiness for the next wave of scientific innovation, and to encourage

more companies to use AI and machine learning in Singapore. It mainly emphasises the use of AI in three key industries: finance, city management solutions and healthcare.[480]

In order to address the ethical challenges arising from AI, a council has been set up to advise the government on the ethical and legal use of AI and data. The Council will include private sector thought leaders in AI from local and international companies as well as representatives from consumer interest.[481] The Singapore Management University will support the work of the council through a five-year research program.[482] This program investigates "ethical, legal, policy and governance issues" based on AI and data use.

In addition, a discussion paper was released by the Personal Data Protection Commission on responsible development and adoption of AI which mentioned two core principles: (i) decisions made by or with the assistance of AI should be explainable, transparent and fair to consumers,

---

[472] See https://thenextweb.com/artificial-intelligence/2018/09/07/california-law-would-make-political-bots-illegal-unless-they-admit-theyre-bots/
[473] Ibid.
[474] U.S. Const. amend. I (1791)
[475] See https://www.nato-pa.int/news/artificial-intelligence-central-all-future-defence-capabilities
[476] See https://www.wired.com/story/how-pentagons-move-to-cloud-landed-in-mud/
[477] Ibid.
[478] See https://dod.defense.gov/News/Article/Article/1356172/project-maven-industry-day-pursues-artificial-intelligence-for-dod-challenges/
[479] See https://www.fedscoop.com/project-maven-artificial-intelligence-google/
[480] See https://euagenda.eu/upload/publications/untitled-128126-ea.pdf
[481] See https://www.opengovasia.com/singapore-announces-initiatives-on-ai-governance-and-ethics/
[482] See https://www.smu.edu.sg/news/2018/06/05/smu-school-law-awarded-significant-research-grant-address-governance-ai-and-data-use

and (ii) AI systems, robots and decisions should be human-centric. The discussion paper was drafted in consultation with key government and industry stakeholders, to support collaborative discussions on the responsible development and adoption of AI.[483]

5.4 International organisations

United Nations

The United Nations (UN) have released a document which discusses the *United Nations Activities on Artificial Intelligence (AI)* by the different UN agencies. [484] [485]

For instance, the document includes a summary of the *International Labour Organisation's* (ILO) report on the impact of artificial intelligence on jobs, inequality and the change of productivity due to technological change. In addition to that, the ILO also works on a study of potential of AI tools to measure skills demand and development in developing countries. *The United Nations Conference on Trade and Development* (UNCTAD) launched a *Technology and Innovation Report* in 2018, which discusses the potential of frontier technology as an opportunity for developing countries to accelerate progress towards their Sustainable Development Goals (SDGs).[486] The report emphasises the positive impact Artificial Intelligence can have to reach these goals, but at the same time warns of its potential negative effects on vulnerable groups, because of wealth and knowledge-concentrating tendencies.

The *United Nations Educational, Scientific and Cultural Organization* (UNESCO) is planning an event in January 2019 in Paris, in which the potential transformations of societies in the UNESCO's fields of competence will be discussed.[487] The organisation plans to identify the potential risks and benefits of AI transformation processes regarding ethical, social and human rights implications in particular. Furthermore, the *United Nations High Commissioner for Refugees* (UNHCR) initiated a predictive analysis in 2017 to forecast population movements in the Horn of Africa, starting with Somalia.[488] This so-called *Project Jetson* applies machine-learning and automation processes and aims to expand this technology to regions beyond Somalia soon.

The *United Nations Office for Disaster Risk Reduction* (UNISDR) is applying AI technologies for the development of their forthcoming *Global Risk Assessment Framework* (GRAF).[489] This framework aims at understanding future risk conditions on Earth to manage uncertainties and improve risk-informed decision-making at all scales and across relevant time periods.

The *United Nations Special Rapporteur on Extreme Poverty and Human Rights* released a statement upon a visit to the United Kingdom focussing on the challenges of algorithmic systems in enabling access to government services. This report criticised the 'the limits of an ethics frame' to the AI governance institutions being established in, for example, the United

[483] See https://www.pdpc.gov.sg/Resources/Discussion-Paper-on-AI-and-Personal-Data
[484] Note that the section on the United Nations only provides a non-exhaustive list of projects within some of the United Nations' more than 30 agencies.
[485] United Nations Activities on Artificial Intelligence (2018) available at: https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2018-1-PDF-E.pdf
[486] Ibid. p. 19.
[487] Ibid. p. 28.
[488] Ibid. p. 34.
[489] Ibid. p. 44.

Kingdom, calling for '[g]overnment use of automation, with its potential to severely restrict the rights of individuals, [to] be bound by the rule of law and not just an ethical code'.[490]

Lastly, the *World Health Organization* (WHO) aims at making health services across the world more accessible and more effective by making data collection and triage more efficient through AI.[491] According to the organisation, AI can also be applied to improve disease surveillance and prevent outbreaks, by allowing health authorities to closely monitor emerging infectious diseases and enable them to react more quickly with suitable containment efforts.

The '**AI for Good' series is the leading United Nations (UN) platform for dialogue on AI**. It is a yearly summit, initiated in 2017, aiming to "identify practical applications of AI and supporting strategies to improve the quality and sustainability of life on our planet"[492]. The exercise connects AI innovators with public and private-sector decision-makers with the objective of building collaboration. In 2018, the Global Summit focused on "inclusive development of AI technologies and equitable access to their benefits", covering issues such as the explainability, transparency and robustness of AI algorithms.[493]

The International Telecommunication Union's (ITU) specific workstreams in this context are explained in more depth in its report *Artificial Intelligence for global good*[494], which provides a comprehensive account of the UN's positioning on the potential impacts of AI and machine learning on society.

In addition to ITU's 'AI for Good' series, the UN, through the United Nations Interregional Crime and Justice Research Institute (UNICRI), has established the **Centre on Artificial Intelligence and Robotics**[495] in The Hague. The Centre is mostly focussed on the impacts of using AI in areas such as crime prevention, criminal justice, law enforcement and national security, as well as the legal, social and ethical concerns involved. The Centre is thus "dedicated to understanding and addressing the risks and benefits of AI and robotics from the perspective of crime and security through awareness-raising, education, exchange of information, and harmonization of stakeholders"[496]. As part of this initiative, UNICRI has developed a network of stakeholders with whom it collaborates, including the International Criminal Police Organization (INTERPOL), the International Telecommunications Union (ITU), the Institute of Electrical and Electronics Engineers (IEEE), the Foundation for Responsible Robotics, the World Economic Forum, Centre for Future Intelligence (CFI) and others[497].

Organisation for Economic Co-operation and Development (OECD)

Following its extensive work on AI through the Going Digital project[498], the OECD created a multidisciplinary expert group, entitled AIGO and made up of stakeholders from think tanks,

[490] https://www.ohchr.org/Documents/Issues/Poverty/EOM_GB_16Nov2018.pdf
[491] Opt. Cit. United Nations Activities on Artificial Intelligence, p. 56.
[492] See https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx
[493] See https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx
[494] See https://www.itu.int/en/itunews/Documents/2018/2018-01/2018_ITUNews01-en.pdf
[495] See http://www.unicri.it/topics/ai_robotics/centre/
[496] Ibid.
[497] Ibid.
[498] See http://www.oecd.org/going-digital/ai/

civil society and labour associations.[499] AIGO aims to provide guidance on scoping for artificial intelligence in society[500]. The main objective of the expert group is to support governments, business, labour and the public to "maximise the benefits of AI and minimise its risks". Practically, the group will work towards this goal by leveraging its wide stakeholder base to **co-develop principles** to "keep countries competitive, guide the ethical progress of AI, and share knowledge with the broader world"[501]. In particular, the OECD's work references specific focus on: ensuring AI does not exacerbate inequality; mitigation of biases; and ensuring the safety, security, transparency and accountability of AI.

A key achievement of the OECD's focus on AI was the October 2017 OECD conference on AI: Intelligent Machines – Smart Policies. This conference brought together more than 300 stakeholders from industry, government and civil society. Furthermore, in March 2018, the G7 Innovation Ministers agreed to "facilitate multistakeholder dialogue and collaboration on artificial intelligence to inform future policy discussions by G7 governments, supported by the OECD in its multistakeholder convener role".[502]

Ongoing work by the OECD includes:

- mapping the **economic and social impacts of AI technologies and applications** and their policy implications;
- planning to set up a ***Policy Observatory on AI***[503], which is to be launched in 2019. The purpose of the observatory will be to provide insight on public policies to ensure the beneficial use of AI across and in support of governments, as well as a vehicle for "engaging different stakeholder groups"[504].

Council of Europe

The Council of Europe has set out a Committee of experts[505] on human rights dimensions of automated data processing and different forms of artificial intelligence. Work of the expert group has started in 2018, and findings are expected to be presented in 2019.

5.5 Intermediate findings

A clear conclusion on the **policy responses to the challenges facing algorithmic decision-making is that, across the globe, the majority of initiatives are very recent or still in development.** Additionally, there are **limited concrete legislative or regulatory initiatives being implemented across the globe, but public attention as to the risks, ethical implications and fundamental rights concerns raised by automation are emerging in parallel to, or as part of, strategies for competitiveness in AI development and uptake**.

---

[499] A full list of participating stakeholders can be accessed at http://www.oecd.org/going-digital/ai/oecd-aigo-membership-list.pdf

[500] See http://www.oecd.org/going-digital/ai/oecd-creates-expert-group-to-foster-trust-in-artificial-intelligence.htm

[501] See http://www.oecd.org/going-digital/ai/oecd-creates-expert-group-to-foster-trust-in-artificial-intelligence.htm

[502] See http://www.oecd.org/going-digital/ai/oecd-initiatives-on-ai.htm

[503] Ibid.

[504] Ibid.

[505] https://www.coe.int/en/web/freedom-expression/msi-aut

This is not to say however that algorithmic decision-making operates in a deregulated environment. The regulatory framework applied is generally technology-neutral, and rules applicable in specific sectors are not legally circumvented by the use of automated tools, as opposed to human decisions. Legal frameworks such as fundamental rights, national laws on non-discrimination, consumer protection legislation, competition law, safety standards still apply. Where concrete legislation has been enacted in the EU, the prominent examples relate primarily to the protection of personal data, primarily the EU's GDPR and national laws supporting the application of the Regulation. Jurisdictions such as the US have not yet implemented a comparable and comprehensive piece of legislation regulating personal rights. This might change to a certain extent with the introduction of the Future of AI bill, which includes more provisions on the appropriate use of algorithm-based decision-making. On the state level, the focus mainly is set on the prohibition of the use of non-disclosed AI bots (deriving from experiences of Russian AI bots intervening in the US Presidential election 2016) and the regulation of the use of automated decision-making by public administration.

The concept of algorithmic accountability discussed in section 3.3 should also be contextualized in the light of the policy initiatives. Indeed, the debate on accountability stems mainly from the United States, and while the societal aspects of the debate are very relevant and interesting, they reflect a situation where the legal context is very different than in the EU. The introduction of the GDPR means that a large part of the debate on accountability for processing of personal data is not as such relevant in the EU context. However, the practical application of the GDPR, methodological concerns as to AI explainability, methods for risk and impact assessment, and practical governance questions are more pertinent to the EU debate.

A few examples of AI-specific legislation have been identified, but the underlying question remains as to the need for assessing rule-making targeting a technology, or rather specific policy and regulatory environments adapted to the areas of application of the technology, and the consequent risks and stakes in each instance.

More commonly, however, the initiatives identified are softer in nature. These initiatives also reflect the aim of harnessing the potential of AI through the development of wide-reaching industrial and research strategies. Prominent types of initiatives implemented globally include:

- Development of **strategies on the use of AI and algorithmic decision-making**, with examples including France's AI for Humanity Strategy, which focuses on driving AI research, training and industry in France alongside the development of an ethical framework for AI to ensure, in particular, transparency, explainability and fairness. Another example is the Indian National AI Strategy and the €3bn AI strategy issued by Germany in November 2018, which aims at making the country a frontrunner in the second AI wave, while maintaining strong ethical principals.[506] Similar to this are the numerous White Papers and reports developed, including the German White Paper on AI, the Visegrád position paper on AI and the Finnish Age of AI report.
- Establishment of **expert groups and guidance bodies** with examples including the Group of Experts and "Sages" established in Spain in 2018, the Italian AI Task Force and the German Enquete Commission. Considering the former example, this group has

[506] See https://www.politico.eu/article/germanys-plan-to-become-an-ai-powerhouse/

been tasked with guiding on the ethics of AI and Big Data through an examination of the social, juridicial and ethical implications of AI.

6. Next steps and further research

This report represents an evolving account of the ongoing academic debate around the impacts of algorithmic decision-making, as well as a review of relevant initiatives within industry and civil society, and policy initiatives and approaches adopted by several EU and third countries.

As previously highlighted, the intention is for the report to undergo a peer-review process and to be made available online to allow for comments by interested stakeholders. Through the website and our social media channels **algo:aware** will engage with individuals from academia, government, civil society and industry to ensure that we capture the perspectives from each stakeholder group. This is a fundamental part of the development of the *State-of-the-Art* report to ensure that we have captured the latest thoughts and analysis of subject matter experts and have mapped a comprehensive list of policy initiatives -and their expected impact – in the field.

The analysis within the report already indicate some underexplored areas and directions of future research. These areas will be further investigated through a series of case studies which will look to provide an in-depth contextualisation and analysis from fairness, accountability, transparency and robustness standpoints.

Bibliography

Accenture. “An Ethical Framework for Responsible AI and Robotics.” Accessed October 23, 2018. https://www.accenture.com/gb-en/company-responsible-ai-robotics.

AI Now. “LITIGATING ALGORITHMS: CHALLENGING GOVERNMENT USE OF ALGORITHMIC DECISION SYSTEMS,” 2018.

Anderson, CW. “Towards a Sociology of Computational and Algorithmic Journalism.” *New Media & Society* 15, no. 7 (November 1, 2013): 1005–21. https://doi.org/10.1177/1461444812465137.

Andrews, Robert, Joachim Diederich, and Alan B. Tickle. “Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks.” *Knowledge-Based Systems*, Knowledge-based neural networks, 8, no. 6 (December 1, 1995): 373–89. https://doi.org/10.1016/0950-7051(96)81920-4.

Ausloos, Jef, and Pierre Dewitte. “Shattering One-Way Mirrors – Data Subject Access Rights in Practice.” *International Data Privacy Law* 8, no. 1 (February 1, 2018): 4–28. https://doi.org/10.1093/idpl/ipy001.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*, 2018.

Barocas, Solon, and Helen Nissenbaum. “Big Data’s End Run around Anonymity and Consent.” In *Privacy, Big Data, and the Public Good Frameworks for Engagement*, n.d. https://www.cambridge.org/core/books/privacy-big-data-and-the-public-good/big-datas-end-run-around-anonymity-and-consent/0BAA038A4550C729DAA24DFC7D69946C.

Barocas, Solon, and Andrew D. Selbst. “Big Data’s Disparate Impact.” *SSRN Electronic Journal*, 2016. https://doi.org/10.2139/ssrn.2477899.

Bauer, Johannes, and Michael Latzer. *Handbook on the Economics of the Internet*. Edward Elgar Publishing, 2016. https://doi.org/10.4337/9780857939852.

Beard, T., George Ford, Thomas Koutsky, and Lawrence Spiwak. “Tort Liability for Software Developers: A Law & Economics Perspective, 27 J. Marshall J. Computer & Info. L. 199 (2009).” *The John Marshall Journal of Information Technology & Privacy Law* 27, no. 2 (January 1, 2009). https://repository.jmls.edu/jitpl/vol27/iss2/1.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. “Fairness in Criminal Justice Risk Assessments: The State of the Art.” *ArXiv:1703.09207 [Stat]*, March 27, 2017. http://arxiv.org/abs/1703.09207.

Bertolini, Andrea, Pericle Salvini, Teresa Pagliai, Annagiulia Morachioli, Giorgia Acerbi, Leopoldo Trieste, Filippo Cavallo, Giuseppe Turchetti, and Paolo Dario. “On Robots and Insurance.” *International Journal of Social Robotics* 8, no. 3 (June 1, 2016): 381–91. https://doi.org/10.1007/s12369-016-0345-z.

Bianchi-Berthouze, Nadia, and Andrea Kleinsmith. “Automatic Recognition of Affective Body Expressions.” *The Oxford Handbook of Affective Computing*, January 1, 2015. https://doi.org/10.1093/oxfordhb/9780199942237.013.025.

Binns, Reuben. “Fairness in Machine Learning: Lessons from Political Philosophy.” *ArXiv:1712.03586 [Cs]*, December 10, 2017. http://arxiv.org/abs/1712.03586.

Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. “‘It’s Reducing a Human Being to a Percentage’: Perceptions of Justice in Algorithmic Decisions.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 377:1–377:14. CHI ’18. New York, NY, USA: ACM, 2018. https://doi.org/10.1145/3173574.3173951.

Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt. “Like Trainer, like Bot? Inheritance of Bias in Algorithmic Content Moderation.” *ArXiv:1707.01477 [Cs]* 10540 (2017): 405–15. https://doi.org/10.1007/978-3-319-67256-4_32.

Borgesius, Frederik J. Zuiderveen. “Personal Data Processing for Behavioural Targeting: Which Legal Basis?” *International Data Privacy Law* 5, no. 3 (August 1, 2015): 163–76. https://doi.org/10.1093/idpl/ipv011.

Borgesius, Frederik J. Zuiderveen, Damian Trilling, Judith Möller, Balázs Bodó, Claes H. de Vreese, and Natali Helberger. “Should We Worry about Filter Bubbles?” *Internet Policy Review*, March 31, 2016. https://policyreview.info/articles/analysis/should-we-worry-about-filter-bubbles.

Bovens, Mark. “Analysing and Assessing Public Accountability. A Conceptual Framework,” n.d.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out*. Accessed September 28, 2018. https://mitpress.mit.edu/books/sorting-things-out.

Brauneis, Robert, and Ellen P. Goodman. “Algorithmic Transparency for the Smart City,” August 2, 2017. https://papers.ssrn.com/abstract=3012499.

Brown, Ian, and Christopher Marsden. “Regulating Code.” The MIT Press. Accessed December 3, 2018. https://mitpress.mit.edu/books/regulating-code.

Brynjolfsson, Erik, and Andrew Mcafee. “The Business of Artificial Intelligence.” Harvard Business Review, July 18, 2017. https://hbr.org/2017/07/the-business-of-artificial-intelligence.

Burri, Mira. “Regulating Code: Good Governance and Better Regulation in the Information Age, by Ian Brown and Christopher T. Marsden.” *International Journal of Law and Information Technology* 22, no. 2 (June 1, 2014): 208–14. https://doi.org/10.1093/ijlit/eat019.

Butler, Alan. “Products Liability and the Internet of (Insecure) Things: Should Manufacturers Be Liable for Damage Caused by Hacked Devices?” 50 (n.d.): 19.

C-236/09 Test-Achats, judgment of 1 March 2011 (n.d.).

Calo, Ryan. “Robotics and the Lessons of Cyberlaw.” *CALIFORNIA LAW REVIEW* 103 (n.d.): 52.

Carlini, Nicholas, Chang Liu, Jernej Kos, Úlfar Erlingsson, and Dawn Song. “The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets.” *ArXiv:1802.08232 [Cs]*, February 22, 2018. http://arxiv.org/abs/1802.08232.

Chen, Le, Ruijun Ma, Anikó Hannák, and Christo Wilson. “Investigating the Impact of Gender on Rank in Resume Search Engines.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18*, 1–14. Montreal QC, Canada: ACM Press, 2018. https://doi.org/10.1145/3173574.3174225.

Chouldechova, Alexandra. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *ArXiv:1610.07524 [Cs, Stat]*, October 24, 2016. http://arxiv.org/abs/1610.07524.

Christiano, Paul. “Human-in-the-Counterfactual-Loop.” AI Alignment, January 21, 2015. https://ai-alignment.com/counterfactual-human-in-the-loop-a7822e36f399.

CNIL. “HOW CAN HUMANS KEEP THE UPPER HAND? The Ethical Matters Raised by Algorithms and Artificial Intelligence,” n.d.

Co, Kenneth T., Luis Muñoz-González, and Emil C. Lupu. “Procedural Noise Adversarial Examples for Black-Box Attacks on Deep Neural Networks,” September 30, 2018. https://arxiv.org/abs/1810.00470.

"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." Accessed September 28, 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN.

Copeland, Eddie. "10 Principles for Public Sector Use of Algorithmic Decision Making." nesta. Accessed October 23, 2018. https://www.nesta.org.uk/blog/10-principles-for-public-sector-use-of-algorithmic-decision-making/.

Courtland, Rachel. "Bias Detectives: The Researchers Striving to Make Algorithms Fair." *Nature*, June 20, 2018.

Cowls, Josh, and Luciano Floridi. "Prolegomena to a White Paper on an Ethical Framework for a Good AI Society." *SSRN Electronic Journal*, 2018. https://doi.org/10.2139/ssrn.3198732.

Crawford, Kate. "The Hidden Biases in Big Data." Harvard Business Review, April 1, 2013. https://hbr.org/2013/04/the-hidden-biases-in-big-data.

Crawford, Kate, and Jason Schultz. "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms." *Boston College Law Review* 55 (2014): 93.

Custers, Bart, Toon Calders, Bart Schermer, and Tal Zarsky, eds. *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Studies in Applied Philosophy, Epistemology and Rational Ethics. Berlin Heidelberg: Springer-Verlag, 2013. //www.springer.com/gb/book/9783642304866.

Datta, Anupam, Shayak Sen, and Yair Zick. "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems." In *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. San Jose, CA: IEEE, 2016. https://doi.org/10.1109/SP.2016.42.

Deibert, Ronald, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain. *Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace*. Vol. 48, 2010. http://choicereviews.org/review/10.5860/CHOICE.48-2125.

Dellarocas, Chrysanthos, Juliana Sutanto, Mihai Calin, and Elia Palme. "Attention Allocation in Information-Rich Environments: The Case of News Aggregators." *Management Science* 62, no. 9 (December 10, 2015): 2543–62. https://doi.org/10.1287/mnsc.2015.2237.

Diab, W. "About JTC 1/SC 42 Artificial Intelligence." *ISO/IEC JTC 1* (blog), May 30, 2018. https://jtc1info.org/jtc1-press-committee-info-about-jtc-1-sc-42/.

Diakopoulos, Nicholas, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, et al. "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML." Accessed October 23, 2018. http://www.fatml.org/resources/principles-for-accountable-algorithms.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. "Fairness Through Awareness." *ArXiv:1104.3913 [Cs]*, April 19, 2011. http://arxiv.org/abs/1104.3913.

Dwyer, Rachel E. "Redlining." In *The Blackwell Encyclopedia of Sociology*. American Cancer Society, 2015. https://doi.org/10.1002/9781405165518.wbeosr035.pub2.

Eckersley, Peter. "How Good Are Google's New AI Ethics Principles?" Electronic Frontier Foundation, June 7, 2018. https://www.eff.org/deeplinks/2018/06/how-good-are-googles-new-ai-ethics-principles.

Edwards, Lilian, and Michael Veale. "Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?," n.d., 15.

Edwards, Lilian, and Michael Veale. “Slave to the Algorithm? Why a ‘right to an Explanation’ Is Probably Not the Remedy You Are Looking For.” Accessed September 28, 2018. https://doi.org/10.31228/osf.io/97upg.

Erlich, Yaniv, Tal Shor, Itsik Pe’er, and Shai Carmi. “Identity Inference of Genomic Data Using Long-Range Familial Searches.” *Science*, October 11, 2018, eaau4832. https://doi.org/10.1126/science.aau4832.

Eslami, Motahhare, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. “Communicating Algorithmic Process in Online Behavioral Advertising.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18*, 1–13. Montreal QC, Canada: ACM Press, 2018. https://doi.org/10.1145/3173574.3174006.

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. “‘I Always Assumed That I Wasn’T Really That Close to [Her]’: Reasoning About Invisible Algorithms in News Feeds.” In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–162. CHI ’15. New York, NY, USA: ACM, 2015. https://doi.org/10.1145/2702123.2702556.

European Commission. “Results of the Public Consultation on the Regulatory Environment for Platforms, Online Intermediaries, Data and Cloud Computing and the Collaborative Economy.” Digital Single Market. Accessed October 22, 2018. https://ec.europa.eu/digital-single-market/en/news/results-public-consultation-regulatory-environment-platforms-online-intermediaries-data-and.

Evans, Richard, and Jim Gao. “DeepMind AI Reduces Google Data Centre Cooling Bill by 40%.” DeepMind. Accessed September 28, 2018. https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/.

Fagan, Craig, and Juan Ortiz Freuler. “White Paper Series | Opportunities and Risks in Emerging Technologies.” Accessed October 23, 2018. https://webfoundation.org/research/white-paper-series-opportunities-and-risks-in-emerging-technologies/.

“Fairness and Machine Learning.” Accessed December 3, 2018. https://fairmlbook.org/.

“Fairness in Platform-to-Business Relations.” Accessed October 22, 2018. https://ec.europa.eu/info/law/better-regulation/initiatives/ares-2017-5222469_en.

Farhan, Yue, Morillo, Ware, Lu, Bi, Kamath, Russell, Bamis, and Wang. “Behavior vs. Introspection: Refining Prediction of Clinical Depression via Smartphone Sensing Data.” In *2016 IEEE Wireless Health (WH)*, 1–8, 2016. https://doi.org/10.1109/WH.2016.7764553.

Forsythe, D. E. “New Bottles, Old Wine: Hidden Cultural Assumptions in a Computerized Explanation System for Migraine Sufferers.” *Medical Anthropology Quarterly* 10, no. 4 (December 1996): 551–74.

Forsythe, Diana E. “Using Ethnography in the Design of an Explanation System.” *Expert Systems with Applications*, Explanation: The Way Forward, 8, no. 4 (January 1, 1995): 403–17. https://doi.org/10.1016/0957-4174(94)E0032-P.

Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. “Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures.” In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS ’15*, 1322–33. Denver, Colorado, USA: ACM Press, 2015. https://doi.org/10.1145/2810103.2813677.

Gama, João, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. “A Survey on Concept Drift Adaptation.” In *ACM Computing Surveys*, 46:1–37, 2014. https://doi.org/10.1145/2523813.

Garfinkel, Robert, Ram D. Gopal, Bhavik K. Pathak, Rajkumar Venkatesan, and Fang Yin. “Empirical Analysis of the Business Value of Recommender Systems.” *SSRN Electronic Journal*, 2006. https://doi.org/10.2139/ssrn.958770.

Glover, Eric J., Steve Lawrence, William P. Birmingham, and C. Lee Giles. “Architecture of a Metasearch Engine That Supports User Information Needs.” In *Proceedings of the Eighth International Conference on Information and Knowledge Management - CIKM ’99*, 210–16. Kansas City, Missouri, United States: ACM Press, 1999. https://doi.org/10.1145/319950.319980.

Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. “Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators.” *Journal of the American Medical Informatics Association: JAMIA* 19, no. 1 (February 2012): 121–27. https://doi.org/10.1136/amiajnl-2011-000089.

Grgic-Hlaca, Nina, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. “Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning,” n.d., 10.

Gröndahl, Tommi, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. “All You Need Is ‘Love’: Evading Hate-Speech Detection,” August 28, 2018. https://arxiv.org/abs/1808.09115.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. “A Survey Of Methods For Explaining Black Box Models.” *ArXiv:1802.01933 [Cs]*, February 6, 2018. http://arxiv.org/abs/1802.01933.

Gunes, H., and M. Piccardi. “Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, no. 1 (February 2009): 64–84. https://doi.org/10.1109/TSMCB.2008.927269.

Gunning, David. “Explainable Artificial Intelligence.” Accessed October 23, 2018. https://www.darpa.mil/program/explainable-artificial-intelligence.

Hardt, Moritz, Eric Price, and Nathan Srebro. “Equality of Opportunity in Supervised Learning.” *ArXiv:1610.02413 [Cs]*, October 7, 2016. http://arxiv.org/abs/1610.02413.

Healey, Jennifer. “Physiological Sensing of Emotion.” In *The Oxford Handbook of Affective Computing*, 2015. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199942237.001.0001/oxfordhb-9780199942237-e-023.

Hildebrandt, Mireille. “The Dawn of a Critical Transparency Right for the Profiling Era.” *Stand Alone*, 2012, 41–56. https://doi.org/10.3233/978-1-61499-057-4-41.

Hildebrandt, Mireille, and Serge Gutwirth, eds. *Profiling the European Citizen: Cross-Disciplinary Perspectives*. Springer Netherlands, 2008. //www.springer.com/gb/book/9781402069130.

IBM. “IBM’S Principles for Data Trust and Transparency.” THINKPolicy, May 30, 2018. https://www.ibm.com/blogs/policy/trust-principles/.

IEEE. “ETHICALLY ALIGNED DESIGN.” The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, n.d.

IEEE. “The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems - Executive Committee Descriptions & Members.” The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, n.d.

Issenberg, Sasha. “The Victory Lab.” PenguinRandomhouse.com. Accessed December 3, 2018. https://www.penguinrandomhouse.com/books/215192/the-victory-lab-by-sasha-issenberg/9780307954800.

ITI. “AI Policy Principles.” Information Technology Industry Council, n.d.

ITU News. “Artificial Intelligence for Global Good.” International Telecommunication Union, n.d.

J. Russell, S, and Peter Norvig. *Artificial Intelligence, A Modern Approach. Second Edition*, 2003.

Kamiran, Faisal, Toon Calders, and Mykola Pechenizkiy. “Techniques for Discrimination-Free Predictive Models.” In *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, edited by Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky, 223–39. Studies in Applied Philosophy, Epistemology and Rational Ethics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. https://doi.org/10.1007/978-3-642-30487-3_12.

Kay, Judy. “Scrutable Adaptation: Because We Can and Must.” In *Adaptive Hypermedia and Adaptive Web-Based Systems*, edited by Vincent P. Wade, Helen Ashman, and Barry Smyth, 11–19. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006.

Kemper, Jakko, and Daan Kolkman. “Transparent to Whom? No Algorithmic Accountability without a Critical Audience.” *Information, Communication & Society* 0, no. 0 (June 18, 2018): 1–16. https://doi.org/10.1080/1369118X.2018.1477967.

Kilbertus, Niki, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. “Blind Justice: Fairness with Encrypted Sensitive Attributes.” *ArXiv:1806.03281 [Cs, Stat]*, June 8, 2018. http://arxiv.org/abs/1806.03281.

Kleinsmith, A., N. Bianchi-Berthouze, and A. Steed. “Automatic Recognition of Non-Acted Affective Postures.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41, no. 4 (August 2011): 1027–38. https://doi.org/10.1109/TSMCB.2010.2103557.

Knight, Will. “AI Winter Isn’t Coming, Says Baidu’s Andrew Ng.” MIT Technology Review. Accessed September 28, 2018. https://www.technologyreview.com/s/603062/ai-winter-isnt-coming/.

Kohl, Uta. “Google: The Rise and Rise of Online Intermediaries in the Governance of the Internet and beyond (Part 2).” *International Journal of Law and Information Technology* 21, no. 2 (June 1, 2013): 187–234. https://doi.org/10.1093/ijlit/eat004.

Kohl, Uta. “The Rise and Rise of Online Intermediaries in the Governance of the Internet and beyond – Connectivity Intermediaries.” *International Review of Law, Computers & Technology* 26, no. 2–3 (November 1, 2012): 185–210. https://doi.org/10.1080/13600869.2012.698455.

König, René, and Miriam Rasch, eds. *Society of the Query Reader: Reflections on Web Search*. Inc Reader 9. Amsterdam: Inst. of Network Cultures, 2014.

Kroll, Joshua, Joanna Huey, Solon Barocas, Edward Felten, Joel Reidenberg, David Robinson, and Harlan Yu. “Accountable Algorithms.” *University of Pennsylvania Law Review* 165, no. 3 (January 1, 2017): 633.

Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. “Adversarial Examples in the Physical World.” In *ArXiv:1607.02533 [Cs, Stat]*, 2016. http://arxiv.org/abs/1607.02533.

Küsters, Ulrich, B.D. McCullough, and Michael Bell. “Forecasting Software: Past, Present and Future.” *International Journal of Forecasting* 22, no. 3 (January 2006): 599–615. https://doi.org/10.1016/j.ijforecast.2006.03.004.

Langford, Andrew. “GMonopoly: Does Search Bias Warrant Antitrust or Regulatory Intervention?” *INDIANA LAW JOURNAL* 88 (n.d.): 35.

Langley, Pat. “The Changing Science of Machine Learning.” *Machine Learning* 82, no. 3 (March 2011): 275–79. https://doi.org/10.1007/s10994-011-5242-y.

Lee, Dave. “Computer Wins Series against Go Master.” *BBC News*, March 12, 2016, sec. Technology. https://www.bbc.com/news/technology-35785875.

Levy, Steven. “Can an Algorithm Write a Better News Story Than a Human Reporter?” *Wired*, April 24, 2012. https://www.wired.com/2012/04/can-an-algorithm-write-a-better-news-story-than-a-human-reporter/.

Lewandowski, Dirk. “Why We Need an Independent Index of the Web.” *ArXiv:1405.2212 [Cs]*, May 9, 2014. http://arxiv.org/abs/1405.2212.

Lim, Brian Y., and Anind K. Dey. “Assessing Demand for Intelligibility in Context-Aware Applications.” In *Proceedings of the 11th International Conference on Ubiquitous Computing - Ubicomp ’09*, 195. Orlando, Florida, USA: ACM Press, 2009. https://doi.org/10.1145/1620545.1620576.

Lim, Brian Y., Anind K. Dey, and Daniel Avrahami. “Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2119–2128. CHI ’09. New York, NY, USA: ACM, 2009. https://doi.org/10.1145/1518701.1519023.

LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique, 2016-1321 § (2016).

Luger, Ewa, and Tom Rodden. “An Informed View on Consent for UbiComp.” In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 529–538. UbiComp ’13. New York, NY, USA: ACM, 2013. https://doi.org/10.1145/2493432.2493446.

Machine Intelligence Garage. “Ethics Framework - Responsible AI.” MI Garage. Accessed October 23, 2018. https://www.migarage.ai/ethics-framework/.

Mahieu, Rene, Hadi Asghari, and Michel van Eeten. “Collectively Exercising the Right of Access: Individual Effort, Societal Effect.” Rochester, NY: Social Science Research Network, 2017. https://papers.ssrn.com/abstract=3107292.

Marquess, Kate. “Redline May Be Going Online: Dot-Com Delivery Service Faces Same Complaints as Brick-and-Mortar Peers.” *ABA Journal* 86, no. 8 (2000): 80–95.

McQuillan, Dan. “People’s Councils for Ethical Machine Learning.” *Social Media + Society* 4, no. 2 (April 1, 2018): 2056305118768303. https://doi.org/10.1177/2056305118768303.

Milli, Smitha, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. “Model Reconstruction from Model Explanations.” *ArXiv:1807.05185 [Cs, Stat]*, July 13, 2018. http://arxiv.org/abs/1807.05185.

Mitchell, Thomas M. *Machine Learning*. 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

Moffat, Viva R. “REGULATING SEARCH” 22 (n.d.): 40.

Montavon, Grégoire, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. “Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition.” *Pattern Recognition*, May 2017. https://doi.org/10.14279/depositonce-7011.

Morisy, Michael. “PayPal Practices Defense with Deep Learning.” MIT Technology Review. Accessed October 22, 2018. https://www.technologyreview.com/s/545631/how-paypal-boosts-security-with-artificial-intelligence/.

Nissenbaum, Helen. “Accountability in a Computerized Society.” *Science and Engineering Ethics* 2, no. 1 (March 1, 1996): 25–42. https://doi.org/10.1007/BF02639315.

Nissenbaum, Helen. “Computing and Accountability.” *Commun. ACM* 37, no. 1 (January 1994): 72–80. https://doi.org/10.1145/175222.175228.

Oswald, Marion. “Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power.” *Phil. Trans. R. Soc. A* 376, no. 2128 (September 13, 2018): 20170359. https://doi.org/10.1098/rsta.2017.0359.

Otterlo, Martijn van, and Marco Wiering. “Reinforcement Learning and Markov Decision Processes.” In *Reinforcement Learning: State-of-the-Art*, edited by Marco Wiering and Martijn van Otterlo, 3–42. Adaptation, Learning, and Optimization. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. https://doi.org/10.1007/978-3-642-27645-3_1.

Overdorf, Rebekah, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, and Seda Gürses. “POTs: Protective Optimization Technologies.” *ArXiv:1806.02711 [Cs]*, June 7, 2018. http://arxiv.org/abs/1806.02711.

Pasquale, Frank. “Restoring Transparency to Automated Authority” 9 (n.d.): 22.

Pearl, Judea, and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*, 2018. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1592572.

Pinto, Diane. “Ethical Principles for Artificial Intelligence and Data Analytics,” n.d.

ProPublica. “Machine Bias — ProPublica.” Accessed December 3, 2018. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Rathenau Instituut. “Human rights in the robot age.” Accessed October 22, 2018. https://www.rathenau.nl/en/digital-society/human-rights-robot-age.

Reed, Chris, Elizabeth Kennedy, and Sara Silva. “Responsibility, Autonomy and Accountability: Legal Liability for Machine Learning,” October 17, 2016. https://papers.ssrn.com/abstract=2853462.

“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).” Accessed September 27, 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. “Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability,” n.d.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “Model-Agnostic Interpretability of Machine Learning.” *ArXiv:1606.05386 [Cs, Stat]*, June 16, 2016. http://arxiv.org/abs/1606.05386.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” In *ArXiv:1602.04938 [Cs, Stat]*, 2016. http://arxiv.org/abs/1602.04938.

Rosenblat, Alex, Karen E. C. Levy, Solon Barocas, and Tim Hwang. “Discriminating Tastes: Uber’s Customer Ratings as Vehicles for Workplace Discrimination.” *Policy & Internet* 9, no. 3 (September 1, 2017): 256–79. https://doi.org/10.1002/poi3.153.

Ruan, Sherry, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. “Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones.” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, no. 4 (January 2018): 159:1–159:23. https://doi.org/10.1145/3161187.

Schmon, Christoph. “REVIEW OF PRODUCT LIABILITY RULES,” n.d., 11.

Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. “Hidden Technical Debt in Machine Learning Systems.” In *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, 2503–2511. Curran Associates, Inc., 2015. http://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf.

Selbst, Andrew D., and Solon Barocas. “The Intuitive Appeal of Explainable Machines.” *SSRN Electronic Journal*, 2018. https://doi.org/10.2139/ssrn.3126971.

Sharif, Mahmood, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. “Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition.” In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS’16*, 1528–40. Vienna, Austria: ACM Press, 2016. https://doi.org/10.1145/2976749.2978392.

Shokri, R., M. Stronati, C. Song, and V. Shmatikov. “Membership Inference Attacks Against Machine Learning Models.” In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18, 2017. https://doi.org/10.1109/SP.2017.41.

Singh, Jatinder, Jennifer Cobbe, and Chris Norval. “Decision Provenance: Capturing Data Flow for Accountable Systems.” *ArXiv:1804.05741 [Cs]*, April 16, 2018. http://arxiv.org/abs/1804.05741.

Skitka, LINDA J., KATHLEEN L. Mosier, and MARK Burdick. “Does Automation Bias Decision-Making?” *International Journal of Human-Computer Studies* 51, no. 5 (November 1, 1999): 991–1006. https://doi.org/10.1006/ijhc.1999.0252.

Smith, Brad, and Harry Shum. “The Future Computed: Artificial Intelligence and Its Role in Society.” *The Official Microsoft Blog* (blog), January 18, 2018. https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/.

Sridhar, Vinay, Sriram Subramanian, Dulcardo Arteaga, Swaminathan Sundararaman, Drew Roselli, and Nisha Talagala. “Model Governance: Reducing the Anarchy of Production ML,” 8, n.d.

Steinbrecher, Sandra. “Design Options for Privacy-Respecting Reputation Systems within Centralised Internet Communities.” In *Security and Privacy in Dynamic Environments*, edited by Simone Fischer-Hübner, Kai Rannenberg, Louise Yngström, and Stefan Lindskog, 123–34. IFIP International Federation for Information Processing. Springer US, 2006.

Steiner, Christopher. *Automate This: How Algorithms Took Over Our Markets, Our Jobs, and the World*. Penguin, 2012.

Stepanek, Marcia. “Weblining.” *Bloomberg.Com*, April 3, 2000. https://www.bloomberg.com/news/articles/2000-04-02/weblining.

Taylor, Linnet, Luciano Floridi, and Bart van der Sloot, eds. *Group Privacy: New Challenges of Data Technologies*. Philosophical Studies Series. Springer International Publishing, 2017. //www.springer.com/gp/book/9783319466064.

Tickle, A.B., R. Andrews, M. Golea, and J. Diederich. “The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded within Trained Artificial Neural Networks.” *IEEE Transactions on Neural Networks* 9, no. 6 (November 1998): 1057–68. https://doi.org/10.1109/72.728352.

Tintarev, Nava. “Explaining Recommendations.” In *User Modeling 2007*, edited by Cristina Conati, Kathleen McCoy, and Georgios Paliouras, 470–74. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007.

Tintarev, Nava, and Judith Masthoff. “Explaining Recommendations: Design and Evaluation.” In *Recommender Systems Handbook*, edited by Francesco Ricci, Lior Rokach, and Bracha Shapira, 353–82. Boston, MA: Springer US, 2015. https://doi.org/10.1007/978-1-4899-7637-6_10.

Tjong Tjin Tai, Eric. “Liability for (Semi)Autonomous Systems: Robots and Algorithms,” April 13, 2018. https://papers.ssrn.com/abstract=3161962.

Tramer, Florian, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. “Stealing Machine Learning Models via Prediction APIs,” 19, n.d.

Tranberg, Pernille. “Experts On The Pros & Cons of Algorithms - Dataethical Thinkdotank.” Accessed October 22, 2018. https://dataethics.eu/en/prosconsai/.

Ustun, Berk, and Cynthia Rudin. “Supersparse Linear Integer Models for Optimized Medical Scoring Systems.” *Machine Learning* 102, no. 3 (March 2016): 349–91. https://doi.org/10.1007/s10994-015-5528-6.

Van Kleek, M., W. Seymour, M. Veale, R. Binns, and N. Shadbolt. “The Need for Sensemaking in Networked Privacy and Algorithmic Responsibility.” In *Sensemaking in a Senseless World: Workshop at ACM CHI’18, 22 April 2018, Montréal, Canada*, 2018. http://discovery.ucl.ac.uk/10046886/.

Veale, Michael, and Reuben Binns. “Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data.” *Big Data & Society* 4, no. 2 (December 2017): 205395171774353. https://doi.org/10.1177/2053951717743530.

Veale, Michael, Reuben Binns, and Lilian Edwards. “Algorithms That Remember: Model Inversion Attacks and Data Protection Law.” *ArXiv:1807.04644 [Cs]*, July 12, 2018. https://doi.org/10.1098/rsta.2018.0083.

Veale, Michael, and Lilian Edwards. “Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling.” *Computer Law & Security Review* 34, no. 2 (April 1, 2018): 398–404. https://doi.org/10.1016/j.clsr.2017.12.002.

Veale, Michael, Max Van Kleek, and Reuben Binns. “Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making.” *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18*, 2018, 1–14. https://doi.org/10.1145/3173574.3174014.

Vedder, Anton. “KDD: The Challenge to Individualism.” *Ethics and Information Technology* 1, no. 4 (December 1, 1999): 275–81. https://doi.org/10.1023/A:1010016102284.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.” *SSRN Electronic Journal*, 2017. https://doi.org/10.2139/ssrn.3063289.

Wallace, Julian. “Modelling Contemporary Gatekeeping: The Rise of Individuals, Algorithms and Platforms in Digital News Dissemination.” *Digital Journalism* 6, no. 3 (March 16, 2018): 274–93. https://doi.org/10.1080/21670811.2017.1343648.

Weiser, Marc. “The World Is Not a Desktop.” *Interactions* 1, no. 1 (January 1994): 7–8. https://doi.org/10.1145/174800.174801.

“Wells Fargo Yanks ‘Community Calculator’ Service after ACORN Lawsuit.” Accessed September 28, 2018. https://perma.cc/XG79-9P74.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” September 26, 2016. https://arxiv.org/abs/1609.08144.

Yeung, Karen. “‘Hypernudge’: Big Data as a Mode of Regulation by Design.” *Information, Communication & Society* 20, no. 1 (January 2, 2017): 118–36. https://doi.org/10.1080/1369118X.2016.1186713.

Zeleznikow, J. “The Split-up Project: Induction, Context and Knowledge Discovery in Law.” *Law, Probability and Risk* 3, no. 2 (June 1, 2004): 147–68. https://doi.org/10.1093/lpr/3.2.147.

Zeleznikow, John, and Andrew Stranieri. “The Split-up System: Integrating Neural Networks and Rule-Based Reasoning in the Legal Domain.” In *Proceedings of the 5th International Conference on Artificial Intelligence and Law*, 185–194. ICAIL ’95. New York, NY, USA: ACM, 1995. https://doi.org/10.1145/222092.222235.

Zeng, Jiaming, Berk Ustun, and Cynthia Rudin. “Interpretable Classification Models for Recidivism Prediction.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, no. 3 (June 2017): 689–722. https://doi.org/10.1111/rssa.12227.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. “Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?” *Philosophy & Technology*, September 5, 2018. https://doi.org/10.1007/s13347-018-0330-6.

Zhu, Hongwei, Michael D Siegel, and Stuart E Madnick. “Information Aggregation – A Value-Added E-Service,” n.d., 12.

Žliobaitė, Indrė, Mykola Pechenizkiy, and João Gama. “An Overview of Concept Drift Applications.” In *Big Data Analysis: New Algorithms for a New Society*, edited by Nathalie Japkowicz and Jerzy Stefanowski, 16:91–114. Cham: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-26989-4_4.

Zuiderveen Borgesius, Frederik, and Joost Poort. “Online Price Discrimination and EU Data Privacy Law.” *Journal of Consumer Policy* 40, no. 3 (September 1, 2017): 347–66. https://doi.org/10.1007/s10603-017-9354-z.

algo:aware is procured by the European Commission and delivered by Optimity Advisors.

algo:aware *aims to assess the opportunities and challenges that emerge where algorithmic decisions have a significant bearing on citizens and where they produce societal or economic effects which need public attention.*

www.optimityadvisors.com

www.twitter.com/optimityeurope
www.linkedin.com/company/optimityeurope

Study contact: Quentin Liger – quentin.liger@optimtyadvisors.com