European Parliament

A governance framework for algorithmic accountability and transparency

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 624.262 – April 2019

EN

A governance framework for algorithmic accountability and transparency

Algorithmic systems are increasingly being used as part of decision-making processes in both the public and private sectors, with potentially significant consequences for individuals, organisations and societies as a whole. Algorithmic systems in this context refer to the combination of algorithms, data and the interface process that together determine the outcomes that affect end users. Many types of decisions can be made faster and more efficiently using algorithms. A significant factor in the adoption of algorithmic systems for decision-making is their capacity to process large amounts of varied data sets (i.e. big data), which can be paired with machine learning methods in order to infer statistical models directly from the data. The same properties of scale, complexity and autonomous model inference however are linked to increasing concerns that many of these systems are opaque to the people affected by their use and lack clear explanations for the decisions they make. This lack of transparency risks undermining meaningful scrutiny and accountability, which is a significant concern when these systems are applied as part of decision-making processes that can have a considerable impact on people's human rights (e.g. critical safety decisions in autonomous vehicles; allocation of health and social service resources, etc.).

This study develops policy options for the governance of algorithmic transparency and accountability, based on an analysis of the social, technical and regulatory challenges posed by algorithmic systems. Based on a review and analysis of existing proposals for governance of algorithmic systems, a set of four policy options are proposed, each of which addresses a different aspect of algorithmic transparency and accountability: 1. awareness raising: education, watchdogs and whistleblowers; 2. accountability in public-sector use of algorithmic decision-making; 3. regulatory oversight and legal liability; and 4. global coordination for algorithmic governance.

EPRS | European Parliamentary Research Service

AUTHORS

This study has been written by the following authors at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

Ansgar Koene, main author, University of Nottingham
Chris Clifton, Purdue University
Yohko Hatada, EMLS RI
Helena Webb, Menisha Patel, Caio Machado, Jack LaViolette, University of Oxford
Rashida Richardson, Dillon Reisman, AI Now Institute

ADMINISTRATOR RESPONSIBLE

Mihalis Kritikos, Scientific Foresight Unit (STOA)

To contact the publisher, please e-mail STOA@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in March 2019.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

PE 624.262
ISBN: 978-92-846-4656-2
doi: 10.2861/59990
QA-03-19-162-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

Executive summary

This report presents an analysis of the social, technical and regulatory challenges associated with algorithmic transparency and accountability, including a review of existing proposals for the governance of algorithmic systems and the current state of development of related standards and consideration of the global and human rights dimensions of algorithmic governance.

Motivation

Algorithmic systems are increasingly being used as part of decision-making processes with potentially significant consequences for individuals, organisations and societies as a whole. When used appropriately, with due care and analysis of its impacts on people's lives, algorithmic systems, including artificial intelligence (AI) and machine learning, have great potential to improve human rights and democratic society. In order to achieve this however it is vitally necessary to establish clear governance frameworks for algorithmic transparency and accountability to make sure that the risk and benefits are equitably distributed in a way that does not unduly burden or benefit particular sectors of society. There is growing concern that unless appropriate governance frameworks are put in place, the opacity of algorithmic systems could lead to situations where individuals are negatively impacted because 'the computer says NO', with no recourse to meaningful explanation, a correction mechanism, or a way to ascertain faults that could bring about compensatory processes. As with the governance of any other aspect of society, the extent of algorithmic accountability required should be considered within the context of the good, harm, and risks these systems present.

Background definitions and drivers for algorithmic transparency and accountability

The study presents two 'conceptual landscapes' that explore the conceptual roles and uses of transparency and accountability in the context of algorithmic systems.

The primary role of transparency is identified as a tool to enable accountability. If it is not known what an organisation is doing, it cannot be held accountable and cannot be regulated. Transparency may relate to the data, algorithms, goals, outcomes, compliance, influence and/or usage of automated decision making systems (i.e. algorithmic systems), and will often require different levels of detail for the general public, regulatory staff, third-party forensic analysts and researchers. The degree of transparency of an algorithmic systems often depends on a combination of governance processes and technical properties of the system.

An important difference between transparency and accountability is that accountability is primarily a legal and ethical obligation on an individual or organisation to account for its activities, accept responsibility for them, and to disclose the results in a transparent manner. The challenges for algorithmic accountability arise from: the complex interactions between sub-systems and data sources, which might not all be under the control of the same entity; the impossibility of testing against all possible conditions when there are no formal proofs for the system's performance; difficulties in translating algorithmically derived concepts into human understandable concepts, resulting in incorrect interpretations; information asymmetries arising from algorithmic inferences; accumulation of many small (individually non-significant) algorithmic decisions; difficult to detect injections of adversarial data.

When considering the governance of both transparency and accountability it is important to keep in mind the larger motivating drivers that define what is meant to be achieved. While recognising that fairness is an immensely complex concept with different, sometimes competing, definitions it is nevertheless seen as a fundamental component underpinning responsible systems and it is suggested that algorithmic processes should seek to minimise their potential to be unfair and maximise their potential to be fair. Transparency and accountability provide two important ways in which this can be achieved. Fairness is discussed through the lens of social justice, highlighting the

potential for algorithmic systems to systematically disadvantage, or even discriminate against, different social groups and demographics. A series of real life case studies is used to illustrate how this lack of fairness can arise, before exploring the consequences that lack of fairness can have plus the complexities inherent to trying to achieve fairness in any given societal context. The study describes ways in which lack of fairness in the outcomes of algorithmic systems might be caused by developmental decision-making and design features embedded at different points in the lifecycle of an algorithmic decision making model. A connection is made between the problem of fairness and the tools of transparency and accountability, while highlighting the value of responsible research and innovation (RRI) approaches to pursuing fairness in algorithmic systems.

Technical challenges and solutions

Viewing transparency as 'explaining the steps of the algorithm' is unlikely to lead to an informative outcome. On the one hand, it could result in a description that only captures the general process used to make a decision. At the other extreme would be to provide the complete set of steps taken (e.g. the complete detailed algorithm, or the machine learned model.) While this may enable the outcome to be reconstructed (provided the input data was the same), the complexity is such that even experts may be unable to provide satisfying explanations as to why a particular result was obtained. In order to appreciate the nature of the challenges confronting algorithmic transparency and accountability it is necessary to take into consideration the technical properties of algorithmic decision systems that can give rise to opacity. Key issues are complexity (linked to scale of data, the modularity of algorithms, iterative processing and randomised tiebreaking), the interconnection of decisions, and processes that are learned from data. As a result of these issues, simply releasing the source code of an algorithmic system would often not provide meaningful transparency.

There are also other reasons why simply releasing a model (or the learning algorithm and the data) is often not a feasible solution to transparency: data privacy could be compromised since it may be possible to 'reverse engineer' a model to determine the data used to construct it; continuous, or frequently updated, learning to capture and incorporate new data and changing trends also poses a challenge.

There are, however, technical methods for reducing algorithmic opacity, or extracting explanations for system behaviour despite a lack of transparency. To consider these, it is helpful to divide transparency and explanation into two categories: Understanding the overall system, and understanding a particular outcome. These may require quite different approaches. A key idea to keep in mind is the goal of transparency. Is it to understand how the system works? Or how it behaves?

For understanding the overall system the goal is to obtain a general understanding of the process by which an algorithmic system makes decisions. Approaches include: design/code review; input data analysis; statistical analysis of outcomes; and analysis of sensitivity to inputs. One challenge with these approaches is that they are likely to be difficult or impossible without the direct involvement of system developers. Or at least provision of a 'sandbox' testing environment.

Understanding how a system works is likely of little value for the transparency of individual outcomes. In that case approaches providing explanation become more important. Systems can be designed to provide explanation of the basis of individual outcomes. This can be either a specific design criteria incorporated into the entire system, or accomplished through techniques such as sensitivity analysis.

Meaningful transparency into how outcomes are reached is technically challenging given modern computing systems; regulatory requirements for such transparency may significantly limit the ability to use advanced computing techniques for regulated purposes. Meaningful transparency into the behaviour of computing systems is feasible, and can provide important benefits.

Mechanisms for behavioural transparency may need to be designed into systems, and typically require the participation of the developers or operators of systems.

Technical issues in algorithmic accountability are largely a question if the system behaves according to specifications. Accountability issues such as redress are beyond the technical challenges of the algorithm; these are more a question about the actions implied by the specifications. While accountability for actions taken by algorithmic systems may need to be different than for human actions, those differences are largely governed by the particular application.

Governance frameworks

The review of the governance frameworks of algorithmic transparency and accountability is structured hierarchically. It begins from a high-level perspective on fundamental approaches to technology governance, then provides a detailed consideration of various categories of governance options, and finally reviews specific proposals for governance of algorithmic systems that have been discussed in the existing literature.

At the higher level two aspects are considered, principles vs rules-based approaches and regulation related to algorithms as a single regulatory category or rather as a kind of helper technology that should be regulated as a component of other technologies.

Consideration of principles vs rules-based approaches in the context of technology-related governance reveals that much of the existing literature/practice focuses on risks-oriented principles-based approaches. Methods emphasise maximising the benefits and minimising the risks that arise from the use of the technology by allocating resources in proportion to risks to society, considering both the impacts themselves and the likelihood that they happen, in order to establish appropriate levels of control. One common tool used to support risk-based approaches is an impact assessment.

When considering whether algorithms should be considered as a single regulatory category or, rather, as a component of other kinds of technologies, there are arguments in favour of both, but at the very least, there would need to be strong coordination between agencies when regulating algorithms in order to ensure that lessons learned in developing regulatory solutions for one set of algorithms are readily available to other agencies developing solutions to identical or highly similar algorithms.

At the level of governance mechanisms an analysis is made for each of five governance categories: demand side market solutions; supply side market solutions; companies' self organisation; branches' self-regulation; co-regulation; and state intervention. For each of these a review is made of the current practices related to algorithms or associated technologies, and their likely role in governing algorithmic transparency and accountability. In brief, it is concluded that both demand and supply side market solutions are mostly not effective as regulatory mechanisms for algorithmic transparency or accountability. While there is some movement towards self organisation and self-regulation, much of this appears to be reactive in response to threats to company reputation due to media reports from whistleblowers or investigative journalism. Self-regulation at the level of industry standards setting has started to take shape, but is still in the development stages. Once completed, industry Standards may provide a useful vehicle for co-regulation. At the level of state intervention consideration is given to possible roles for: information measures, e.g. public algorithmic literacy; incentives by funding and taxes, such as strategic investment to increase research into algorithmic methods that are transparent and accountable, as well as technical/infrastructure support for investigative tech-journalism; legislative measures; and a possible role for a regulatory body.

Among legislative measures consideration is given to the role of the General Data Protection Regulation (GDPR), concluding that is not likely to be sufficient. Proposals to maintain 'goal and outcomes' transparency are noted as a means to provide for third-party auditing, while there is also a suggestion based on applying no-fault/strict tort liability with possible modification of liability levels depending on the transparency and criticality of the algorithmic systems.

Possible roles for a regulatory body include: involvement in standards-setting, along the lines of co-regulation with the regulatory body involved in monitoring the use of best-practice standards; a regulator with powers to intervene ranging from 'light touch' nudging of the algorithm designer by means of low-cost incentives to 'hard' intervention by requiring pre-market approval testing.

Existing governance proposals

In the review of existing governance proposals for algorithmic systems attention is given to proposals on: 'a right to reasonable inferences'; a possible role for consumer protection agencies; the establishment of 'an FDA for algorithms'; proposals based on tort liability in combination with algorithm certification by a regulatory agency; an algorithmic impact assessment-based proposal for accountability of algorithmic systems used by public authorities.

Development of industry standards

The development of industry standards relating to algorithmic transparency and accountability methods is important in shaping the potential for branch self-regulation, co-regulation and as reference for the possible setting of regulatory requirements. The review of the current status of standards development in this area however suggests that it will take between 1.5 and four years before the first international standards are completed.

Human rights

While many in the technical and academic communities discussing algorithmic accountability and transparency have framed these in the language of 'ethics', human-rights NGOs and researchers have started to pick up these issues as a matter of human rights. In early 2018 the 'Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems' was published with the aim of drawing attention to the relevant and well-established framework of international human rights law and standards as binding and actionable laws that provide tangible means to protect the human rights of individuals. At implementation level human rights impact assessments provide a framework designed to identify the intended and unintended impact on the enjoyment of human rights, and the State's ability to protect and fulfil them.

Global dimensions of algorithmic governance

As with much of the digital economy, the use of algorithmic systems is characterised by the highly cross-border nature and global reach of the services that are built on these technologies. To successfully govern algorithmic systems therefore requires global dialogue and collaboration across borders and among rich and poor countries to avoid a patchwork of country-specific or regional approaches. In the context of the global dialogue attention is drawn to the tensions arising from the 'winner-takes-all' narrative around the development of artificial intelligence (AI). As a counter point however note is made of the global response to the GDPR, with an increasing number of states enacting GDPR-inspired data privacy legislation, suggesting that the EU may be well positioned to take the lead in establishing a new framework for international coordination for algorithmic accountability and transparency.

Policy options

Based on the review and analysis of current literature regarding algorithmic transparency and accountability, and the successes, failures and challenges of different governance frameworks that have been applied to technological developments (especially in ICT), a set of four policy options is proposed, each of which addresses a different aspect of algorithmic transparency and accountability:

1. awareness raising: education, watchdogs and whistleblowers;
2. accountability in public-sector use of algorithmic decision-making;
3. regulatory oversight and legal liability in the private sector; and
4. the global dimension of algorithmic governance.

Table of contents

1. Introduction

1.1. Motivation

Algorithmic systems are increasingly being used as part of decision-making processes with potentially significant consequences for individuals, organisations and societies as a whole. Because the ways in which these systems reach their 'conclusions' may reflect or amplify existing biases, or may not offer explanations that satisfy our accustomed social and judicial expectations, there is growing concern that the traditional frameworks for implementing transparency and accountability may not suffice as mechanisms of governance.

One of the key areas of concern centres around the opacity of algorithmic systems that can be highly complex and potentially involve machine learning, by which the system behaviour can come to depend not only on design choices at the time of creation but also on the data is it trained on and the input that is being evaluated [1, 2]. Unless appropriate governance frameworks are put in place there is a real risk that situations may arise where individuals are negatively impacted because 'the computer says NO', with no recourse to meaningful explanation, correction mechanism, or way to ascertain faults that could bring about compensatory processes [3, 4].

When linked with pervasive and automated data collection, where the individual is no longer asked to explicitly provide the data that is used by the algorithmic system, it can become difficult or impossible for individuals to identify which data were used to reach particular decision outcomes, and thus impossible to correct faulty data or assumptions, or to even ascertain if an error was made [5, 6, 7].

An important element of concern is the inherent consequence of algorithmic decision-making which implicitly biases social value as being limited to only those things that are measured, since algorithmic decision systems can only take into account those things that are measured and fed in to the system [8, 9]. If blindly applied without democratic and human rights safeguards, this risks driving ever increasing levels of (corporate) surveillance in the name of improved algorithmic decision outcomes [10,11].

Some voices in the academic/research community are also expressing concern that the current high reliance on purely empirical methods (i.e. methods that rely observing responses to sampled test data) for evaluating system performance, with no underlying formal model to support the validity of the system behaviour, is sacrificing scientific rigour for rapid ad-hoc gains with potential long-term costs to the reliability of resulting systems [12]. This could be compared to concerns in school education about teaching methods that produce students who are increasingly good at taking school exams, without truly understanding the material they are being taught. Increased requirements for algorithmic transparency and accountability might contribute to greater focus on scientifically rigorous algorithm evaluation methods, which could lead to significant gains for the future developments in algorithmic systems and Artificial Intelligence [13].

At the same time, algorithmic systems are permeating more and more aspects of our lives -- handwriting analysis, real-time navigation systems, hurricane prediction, medical diagnosis, logistics … -- for the simple reasons that they work better than the systems they are replacing or augmenting. As with the governance of any other aspect of our lives, the extent of algorithmic accountability required should be considered within the context of goods, harms, and risks these systems provide. They require decisions — made by the appropriate stakeholders — that involve often difficult and particular questions of social goals and values, rights, models of fairness, compensatory structures, risk tolerance, etc. [14, 15, 16]

When used appropriately, with due care and analysis of its impacts on people's lives, algorithmic systems, including AI and Machine Learning, have great potential to improve human rights and democratic society [17, 18]. In order to achieve this however it is vitally necessary to establish clear governance frameworks for algorithmic transparency and accountability to make sure that the risk and benefits are equitably distributed in a way that does not unduly burden or benefit particular sectors of society [19]. Such frameworks should not assume that transparency by itself is an ultimate goal, but is a means to support social values; the governance of AI and Machine Learning should therefore also have a tolerance for non-transparent systems when they bring desirable social benefit and carry demonstrably limited and acceptable risks. Transparency is one tool in the governance toolbox. As with any tool, it is crucial that its wielder understand its uses, its limitations [20], and any trade-offs involved in its use.

1.2. Scope

This report reviews possible governance frameworks for accountability and transparency of algorithmic systems, including discussion of the challenges and opportunities associated with their implementation, based on an assessment of the current status of algorithmic system governance and comparison with the governance of other technological systems.

We view the question to be three dimensional: 1. The multiple points in the process and multiple structural elements of algorithmic systems that might be made transparent; 2. The various tools and implements by which algorithmic systems can be governed; 3. The benefits, harms, and risks of governing or leaving ungoverned any of the first two dimensions.

The review includes governance frameworks in terms of legal regulatory mechanism (e.g. impact assessments and auditing requirements for high-impact systems), internal frameworks at the organisational level within public or private sector organisations (e.g. rules governing public procurement of algorithmic systems; adoption of industry standards for algorithmic transparency), and supporting frameworks for promoting third-party investigatory oversight (e.g. support mechanisms for public interest investigative journalism into the use of algorithmic systems).

1.3. Objective

Beyond providing an up-to-date review of current and potential governance frameworks for algorithmic accountability and transparency, the report aims to provide an understanding of the technical challenges and solutions for algorithmic transparency, clarify the relationship between accountability/transparency and fairness, and provide recommendations for an algorithmic impact assessment metric for assessing the degree of regulatory scrutiny of an algorithmic system that would be appropriate/necessary for a particular context of use.

2. Methodology and resources used

The methodology is based on a literature review sourcing primary and secondary scientific literature, including white papers, reports from government inquiries and civil-society investigations into the current state, and proposed future directions, of algorithmic governance. News articles are included as part of the discussion on the role of investigative journalism, as well as illustration of perspectives and concerns in the wider population.

The first step involved a review of technical approaches to algorithmic transparency and algorithmic accountability, including concepts of accountability by design and remedial accountability. The second step included a review of the types and degrees of impact that algorithmic systems have on social justice, fair decision-making and the associated technological and societal need/limits for algorithmic literacy, transparency, oversight and information symmetry. The third step was a review of existing proposals for governance frameworks for algorithmic systems, including relevant sections of EU directive/regulations (e.g. GDPR), national laws/proposals (e.g. French Digital Republic), local government legislation (e.g. New York City law creating a task force to review government use of algorithmic systems) and industry self-regulation (e.g. standards).

2.1. Resources

Literature databases: Google Scholar; Web of Science (key phrase searches for 'algorithmic governance', co-regulation of technology', 'AI ethics regulation', Algorithmic literacy', Algorithmic disclosure'). Key phrase searches were followed by citation and reference based searches.

Types of literature: Primary Literature and Review papers in Computer Science, Law, Social Science; grey literature (Standards documents [e.g. IEEE Standards]; industry reports; media reports [e.g. ProPublica COMPAS]; government reports [e.g. French law for the Digital Republic]);

3. Synthesis of the research work and findings

Here we present an analysis of the social, technical and regulatory challenges associated with algorithmic transparency and accountability, including a review of existing proposals for governance of algorithmic systems, the current state of related Standards development and a considerations of the global and human rights dimensions of algorithmic governance.

We start with definitions for algorithmic transparency and accountability, which are further elaborated through two 'conceptual landscapes' that explore the conceptual roles and uses of transparency and accountability in the context of algorithmic systems. The underlying motivation for a governance framework for algorithmic transparency and accountability is explored further through a review of issues regarding algorithmic fairness. Rounding off this background context setting is an overview of the technical challenges and possible technical solutions that have been identified for algorithmic transparency.

Our review of the governance of algorithmic transparency and accountability is structured hierarchically. We start from a high-level perspective on fundamental approaches to technology governance, then provide a detailed consideration of various categories of governance options, and finally review specific proposals for governance of algorithmic systems that have been discussed in the existing literature.

In the final two subsections we take a more detailed look at the current state of international Standards development related to algorithmic decision-making systems and consider the global and human rights dimensions of algorithmic governance.

3.1. Definitions

The following concise definitions of 'Transparency' and 'Accountability' are provided for reference to clarify the meaning of these key terms as used in this document.

Transparency - Depending on the type and use of an algorithmic decision system, the desire for algorithmic transparency may refer to one, or more of the following aspects: code, logic, model, goals (e.g. optimisation targets), decision variables, or some other aspect that is considered to provide insight into the way the algorithm performs. Algorithmic system transparency can be global, seeking insight into the system behaviour for any kind of input, or local, seeking to explain a specific input - output relationship.

Accountability - 'A set of mechanisms, practices and attributes that sum to a governance structure which involves committing to legal and ethical obligations, policies, procedures and mechanism, explaining and demonstrating ethical implementation to internal and external stakeholders and remedying any failure to act properly' (derived from [21] as used in [22]).

A more detailed consideration of the nuances and meanings associated with the terms is provided in the subsequent 'conceptual landscape' sections.

3.2. Conceptual landscape 1: The uses of transparency

In the context of ensuring responsible development and use of algorithmic systems such that they improve human rights and benefit society, transparency is a tool.

Tools have four properties that are directly relevant to transparency's use as a mechanism of governance of algorithmic systems:

1. A tool is valuable not in itself but because of the goals its serves; a can-opener is only useful if there are cans to be opened.

2. No tool is right for every job. Misusing a tool has costs. Even using it appropriately often requires trade-offs.

3. It cannot be simply assumed that it is a matter of indifference who uses the tool.

So, what is transparency a tool for? What are the limits of its use? What are the costs and trade-offs? What others tools might sometimes be better at the job than transparency?

3.2.1. Transparency of what?

Transparency is implied by the most basic conception of accountability: if we cannot know what an organisation is doing, we cannot hold it accountable, and cannot regulate it.

But the demand for transparency of algorithmic systems goes beyond that simple and assumed sense. We ask for transparent algorithmic systems because they are becoming so central to our lives and economies, and yet some of them use models and algorithms the workings of which are too complex for the human mind to follow.. While the 'black box' metaphor [23] is clearly evocative of this sense of impenetrable mystery of the systems acting upon us, it is important to consider if making the box transparent, so that we can see the gears within, is truly what is needed to satisfy our concerns with these systems. Depending on which aspect of an algorithmic system is in question, that is usually not what the calls for the transparency really aim at [24].

There are seven broad areas of machine learning systems about which transparency might be demanded:

1. **Data.** The transparency of the data used by the algorithmic system -- in particular by machine learning and deep learning algorithms -- can refer to the raw data, to the data's sources, to how the data were preprocessed, to the methods by which it was verified as unbiased and representative (including looking for features that are proxies for information about protected classes), or to the processes by which the data are updated and the system is retrained on them.

2. **Algorithms**. The transparency of the systems' algorithms can refer to testing its output against inputs for which we know the proper output, reducing the variables to the most significant so we can validate them, testing the system with counterfactuals to see if prejudicial data is infecting the output, a third party code review, analysis of how the algorithms work, inspection of internal and external bug reports, or assurance the software development processes are sound.

3. **Goals**. Algorithmic systems can also be transparent about their goals. When a system has multiple goals, this would mean being transparent about their relative priorities. For example, the AI driving autonomous vehicles (AVs) might be aimed at reducing traffic fatalities, lowering the AVs' environmental impact, reducing serious injuries, shortening transit times, avoiding property damage, and providing a comfortable ride. A manufacturer could be required to be transparent about those goals and their priority.

4. **Outcomes**. Manufacturers or operators could be required to be transparent about the outcomes of the deployment of their algorithmic systems, including the internal states of the system (how worn are the brakes of an AV? how much electricity used?), the effects on external systems (how many accidents, or times it's caused another AV to swerve?), and computer-based interactions with other algorithmic systems (what communications with other AVs, what data fed into traffic monitoring systems?).

5. **Compliance**. Manufacturers or operators may be required to be transparent about their overall compliance with whatever transparency requirements have been imposed upon them. In many instances, we may insist that these compliance reports are backed by data that is inspectable by regulators or the general public.

6. **Influence**. Just as the public has an interest in knowing if an article in a newspaper was in fact paid for by an interested party, the public may have an interest in knowing if any element of the AI process was purposefully bent to favour a particular outcome. For example, if a trusted search platform is artificially boosting some results because they were paid to, and if it is not flagging that fact to users, users can be manipulated. Regulators might want to insist that such influence be conspicuously acknowledged.

7. **Usage**. Users may want to know what personal data a system is using, either to personalise outcomes or as data that can train the system to refine it or update it. Knowing what personal data is used, they may then want to control that usage, perhaps to make their personalised results more accurate, or, more urgently, because they feel that usage violates their privacy, even though the data in question may already be a desired part of the system, such as a purchase or search history. There are grey areas here as well: collecting anonymised, highly detailed information about trips made by autonomous vehicles — how often the car brakes or swerves, for example — could be important to optimizing traffic for safety or fuel efficiency. Regulators may face some difficult decisions as well as drawing relatively obvious lines

Note that 'transparency' has different meanings in this categorisation. It can mean: access upon request to the public or authorised people; public posting of information; direct inspection of internal processes; the results of the manufacturer's or operator's tests of the system for accuracy and fairness; delivery of complete subsystems and their data for testing by authorised people, with the results reported to the public or to regulatory bodies; access to computer scientists and managers to explain algorithmic or operational processes.

Transparency is therefore not a single property to be applied blindly to every element of every algorithmic system. It should be applied differently to different systems depending upon the nature of the algorithmic system, the complex circumstances that lead to the need for governance, and the goals of that governance.

3.2.2. By and for whom?

Because transparency often, if not always, has costs and risks, it matters who gets to see what illuminated. When considering regulating transparency the potential viewers include:

- Everyone: fully open access to data, algorithms, outcomes, etc.
- Regulatory staff.
- Third-party forensic analysts whose reports are made public, made selectively public, or kept private.
- Researchers, possibly limited to those affiliated with accredited organisations and/or funding bodies.

3.2.3. Why transparency?

We want systems to be transparent not to satisfy idle curiosity, but to help achieve important social goals related to accountability:

We want to inspect an algorithmic system's ***data*** and ***algorithms*** to:

- Check for bias in the data and algorithms that affects the fairness of the system. (The mechanics, costs, and secondary effects are different when checking data vs. algorithms).
- Check that the system is drawing inferences from relevant and representative data.
- See if we can learn anything from the machine's way of connecting and weighting the data - perhaps there's a meaningful correlation we had not been aware of.
- Look for, and fix, bugs.
- Guard against malicious/adversarial data injection.

We want the hierarchy of ***goals*** and ***outcomes*** to be transparent so:

- It can be debated and possibly regulated.
- Regulators and the public can assess how well an algorithmic system has performed relative to its goals, and compared to the pre-algorithmic systems it may be replacing or supplementing.

We want an organisation's ***compliance*** status to be public so:

- Regulators can hold the organisation accountable in case of failure.
- The public can evaluate the trustworthiness of the organisation, so people can make informed decisions as users about the services offered, and so citizens can become better informed about the benefits, risks, and trade-offs of algorithmic-based services overall.

3.2.4. Fit to purpose

Transparency is not a good in itself, as the struggle for online personal privacy makes evident. Transparency of algorithmic systems is not even clearly a prima facie good, that is — a practice that does not need any special justification, but does require justification for its denial — for transparency of algorithmic systems can have costs associated with it, some of them substantial.

Potential costs include:

- Regulatory bodies have to have staff sufficient to oversee compliance.
- Businesses and other organisations have to create and maintain the processes, code, and legal oversight required by the regulatory bodies.
- Transparency might put justifiable trade secrets at risk.
- Public access to data can flame interest-driven controversies via the untutored or unscrupulous misuse of data [25].
- The requirement for transparency can lead to the use of algorithms that are suboptimal for their purposes, resulting in what can be serious harms when compared with achievable goods.
- Access to data that seems innocuous can lead to breaches of personal privacy by clever and determined hackers.
- Increased transparency of algorithms can make them easier to hack for malicious purposes.

Some of these potential costs can be mitigated by choosing where and how transparency interventions are necessary. For example, rather than providing direct public access to the data being used to train a machine learning system, independent data scientists could examine the data in private and publish the conclusions of their forensic research. Transparency is not an absolute good and thus needs to be negotiated depending on its purpose and the balance of benefits and costs.

Such considerations should include an examination of other tools and remedies to achieve the desired goals. For example, if a system is producing results that replicate, or even amplify, existing biases, allowing the owners of the system to adjust it so that it rights itself might be an acceptable solution. This could be done on the basis of transparency of the system's results, without requiring transparency of the data or algorithms used [26].

Or, if an individual believes that s/he has been discriminated against by a black-box algorithmic system, but there is no evidence of systematic bias, the system might be tested to see if discriminatory factors were determinative in that particular outcome. Such testing might not require transparency. For example, inputting counterfactual data [27] — say, a loan application in which only a factor is changed at a time — can identify the impact of possibly prejudicial data without requiring full transparency.

The costs and benefits of requiring transparency therefore should be weighed based on the benefits and costs, direct and indirect, of using it, and the availability of alternatives.

3.2.5. Conclusion

Transparency is a tool. As with any tool, whether and how it should best be used depends upon:

- The goals of requiring transparency.
- Which elements of the process should be made transparent.
- What type of transparency is most beneficial and least costly.
- Alternative ways of achieving the goal.
- A judgement of the potential trade-off between risk and the benefit an AI system could bring compared with the system it is replacing or augmenting.

Transparency is a tool to be used responsibly, which means accepting that applying it means being sensitive to the complex contexts in which it is used, and the balance of benefits and harms its use inevitably entails.

3.3. Conceptual landscape 2: Accountability

Accountability, like transparency, is ultimately a tool. Accountability serves to ensure responsible development and use of algorithmic systems such that they improve human rights and benefit society [28]. An important difference between transparency and accountability is that accountability is primarily a legal and ethical obligation on an individual or organisation to account for its activities, accept responsibility for them, and to disclose the results in a transparent manner. Transparency, logs of data provenance, code changes and other record keeping are important technical tools but ultimately accountability depends on establishing clear chains of responsibility. Accountability ultimately lies with a (legal) person [29].

In the context of algorithmic systems, the challenges arise from:

- Complex interactions between sub-systems and data sources, some of which might not be under the control of the same entity (e.g. systems relying on data acquired through data brokers who rely on data sources that use algorithmic inference to aggregate over 'similar' data subjects).

- Unexpected outcomes associated with the impossibility of testing against all possible input conditions when there are no methods for generating formal proofs for the system's performance.
- Difficulties in translating algorithmically derived concepts (e.g. clustering algorithm results that segment populations based on large numbers of input variables) into human understandable concepts (e.g. ethnic affiliation) resulting in incorrect interpretations of the meaning of algorithmic results.
- Information asymmetries arising from algorithmic inferences and black box processes that make it all but impossible for data subjects to gage which, potentially false, information might have resulted in a particular algorithmic outcome affecting them (including lack of knowledge that algorithmic processes were even involved).
- Ubiquity of (small) algorithmic decisions which, if systematically biased, may accumulate to have significant impacts on people even though no single decision would have achieved that legal threshold (e.g. impact on personal development due to reinforcement of racial/gender stereotypes by algorithmic recommendations).
- Purposeful injections of adversarial data to fool a system into making errors, often in ways that can be very difficult to detect. [30]

In addition to the basic function of accountability, which is to act as a deterrent to reckless, irresponsible or illegal behaviour on the part of humans deploying/using algorithmic systems,

accountability in algorithmic systems has the potential to generate a self-reflective feedback loop for citizens and society, exposing existing biases and power dynamics [31].

Andrew Tutt [32] summarised the challenge as follows: 'Even if algorithms were programmed with specific attention to well-defined legal norms, it could be extremely difficult to know whether the algorithm behaved according to the legal standard or not in any given circumstance'. This is highlighted with the following example from [33]: 'Algorithms that engage in discrimination offer a good example. Suppose a company used a machine-learning algorithm to screen for promising job candidates. That algorithm could end up discriminating on the basis of race, gender, or sexual orientation—but tracing the discrimination to a problem with the algorithm could be nearly impossible. To be sure, the discrimination could be a result of a bug in the design of the training algorithm, or a typo by the programmer, but it could also be because of a problem with the training data, a by-product of latent society-wide discrimination accidentally channelled into the algorithm, or even no discrimination at all but instead a low-probability event that just happened to be observed.'

In regards to pinpointing human responsibility once illegal or unethical decisions by an algorithmic system have been identified, algorithmic system can again pose challenges since they can be 'sliced-and-diced in a number of ways that many other products are not' [34]. 'A company can sell only an algorithm's code or even give it away. The algorithm could then be copied, modified, customised, and reused or put to use in a variety of applications its initial author never could have imagined. Figuring out how much responsibility the original developer bears when any particular harm arises down the road will be a difficult question. Or consider a second company that sells training data for use in developing one's own learning algorithms, but does not sell any algorithms itself. Depending on the algorithm the customer trains, and the use to which the purchaser wishes to put the data, the data's efficacy could be highly variable, and the responsibility of the data seller could be as well. Or imagine a third company that sells algorithmic services as a package, but the algorithm it offers relies partially or extensively on human interaction when determining its final decisions and outputs (e.g., a stock trading algorithm where a human must confirm all of the proposed trades). Divvying up responsibility between the algorithm and the human is likely to prove complicated.' [32]

Further, as AI systems become more prevalent and more integrated, they will increasing use input from other AI systems, perhaps in highly dynamic ways. For example, as autonomous vehicles are networked, their decisions may result from the AI embedded in scores or hundreds of vehicles, each of which also relies upon independent AI systems providing predictions of micro-weather, traffic congestion, pedestrian flow based on local events, etc. Tracing errors and assessing responsibility may be surpassingly complex — especially since those errors may arise not from individual sources but from the state of transient networks of passing vehicles.

3.4. Algorithmic Fairness: a guiding purpose for transparency and accountability

Fairness is an immensely complex concept with different, sometimes competing, definitions. In contrast to transparency and accountability, we do not suggest that fairness is to be considered as a tool to facilitate best practice in algorithmic systems. Rather we see it as a fundamental component underpinning responsible systems and suggest that algorithmic processes should seek to minimise their potential to be unfair and maximise their potential to be fair. Transparency and accountability provide two important ways in which this can be achieved.

In this section we emphasise the importance of assessing and questioning the fairness of algorithmic systems used for decision-making. We discuss fairness through the lens of social justice and highlight the potential for algorithmic systems to systematically disadvantage, or even discriminate against, different social groups and demographics. We draw on a series of real life case studies to illustrate how this lack of fairness can arise and then go on to explore the consequences lack of fairness can have plus the complexities inherent to trying to achieve fairness in any given societal context. We describe the ways in which lack of fairness in the outcomes of algorithmic systems might be caused by developmental decision-making and design features embedded at different points in the lifecycle of an AI model. We connect the problem of fairness to the tools of transparency and accountability and also highlight the value of responsible research and innovation (RRI) [35] approaches to pursuing fairness in algorithmic systems.

3.4.1. How can we understand algorithmic fairness?

Fairness is an everyday concept that all of us are able to understand at an intuitive level — at least, we usually feel confident about recognizing unfair situations, which, does not mean we have or share a coherent idea of fairness. Fairness turns out to be a multi-faceted, and inherently complex concept. Given this, it is difficult to articulate in a single definition and may also be subject to competing definitions. Fairness reflects the appreciation of a situation based on a set of social values, such as promoting equality in society [36]. The assessment of fairness depends on facts, events, and goals, and therefore has to be understood as situation or task-specific and necessarily addressed within the scope of a practice. Therefore, for the purposes of this report 'fairness' appreciates the social effects of algorithms in sociotechnical structures, while considering how case-specific actions and consequences fit into broad social values. Given the importance of understanding fairness within context, we present a series of relevant case studies of recent controversies regarding the ‘fair’ operation of algorithms in contemporary society. We then use these cases to highlight the core, general issues that require careful attention in discussions of fairness in algorithmic systems.

The concept of fairness in the context of algorithmic implementations appears as a balance between the mutual interests, needs and values of different stakeholders affected by the algorithmic decision. However, these will have varying levels of importance depending on the final application and purpose. As Friedler et al. [37] claim, it is important to have a stated purpose of a given deployment of algorithmic decision-making. The articulated purpose serves both as a benchmark of the algorithmic performance and a legitimizing force, since the relationship between means and ends becomes verifiable.

In this section we are interested in the impact of algorithmic systems on citizens – as individuals and collectively at societal level. Therefore, we also understand fairness within the lens of social justice, as opposed to individual cases in which there is a perceived imbalance of goods or penalties ('Why did she get more cookies than me?), an uneven applications of a rule ('You let him throw the ball out of turn'), a case of discrimination based on irrelevant factors that are not subject to rights claims ('You didn't pick me for the team even though I'm faster than the person you did pick'), etc. Social justice is another complex term with many potential definitions [38, 39, 40, 41, 42]. Broadly speaking, over the course of the 20th century it has come to be understood as referring to a framework that provides the means to achieve a 'fair distribution of societal goods—tangible and intangible' [43]. Discussions of social justice (in academic, policy and public discourses) typically recognise that ensuring a fair distribution is complicated by inherent inequalities in contemporary society; there are various differences of perspective over the extent to which a fair distribution should accommodate for, or attempt to address, such inequalities [44,45]. It has also been pointed out that the values of the goods being distributed are themselves part of complex social systems of practice and values [46]. In the context of this report it is worth noting that new technologies – regardless of their application – are generally dominated by elites such as wealthier social classes and large corporations [47]. In the case of algorithmic systems for decision-making, the design and knowledge of these systems, access to them and ability to influence them are concentrated into the hands of a few. This can be a hindrance to ensuring that they are socially just. Similarly, popular disdain for technocratic domination – real or perceived – can also form a serious impediment.

The following are a number of examples of algorithmic decision-making applications that are most urgently driving the pursuit of a governance structure for algorithmic systems. We ask how useful a tool transparency could be for addressing these issues. What type of transparency? Applied to which parts of the AI system?

Facial recognition systems

Facial recognition technologies identify and/or verify human faces from a digital or video image. Typically, they work via algorithms that identify facial features from the source image and compare them across a dataset. Facial recognition technologies have numerous applications; they are particularly used for security purposes, including in policing and national security activities – including counter terrorism [48, 49]. In recent years, advances in Artificial Intelligence and Machine Learning have increased the capacity and sophistication of these technologies, making them a standard part of consumer goods such as the Apple iPhone which lets users 'sign in' with their faces.

The increasing competency and scalability of this function has given rise to various controversies and concerns. In recent years, several companies—including Microsoft and IBM—have been criticised for rolling out facial recognition software that is more accurate for some demographics than others. Specifically, these systems tend to accurately identify fair-skinned men far more often than they identify darker-skinned women [50]. Similarly, controversy arose when Google's automatic photo-tagging software identified many pictures of African Americans as 'gorilla' or 'monkey' [51]. As discussed later in this section, the cause of these errors is likely to lie in the development of the algorithmic models. The models were presumably trained with datasets of photos of predominantly white people, and thus had not been trained with sufficient data to identify non-white people, particularly women. These inaccuracies have been labelled as (re)producing inequalities by a variety of activists [52, 53]. One high-profile campaigner is Joy Buolamwini, computer scientist at MIT and founder of the Algorithmic Justice League [54]. Her work has prompted multiple companies to release statements addressing criticisms and reform their models [55].

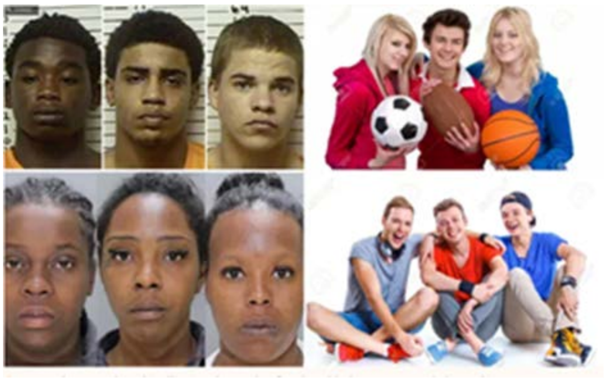
Publicity about the shortcomings of these services can motivate the service providers to do whatever is necessary to fix the problem. Because these are public services, informal outcome transparency is likely enough: users and activists can vet the systems [56]. For services that do not face the public, such as a facial recognition systems used by police forces to identify criminals or dissidents, requiring forensic analysis of outcomes might be called for [57].

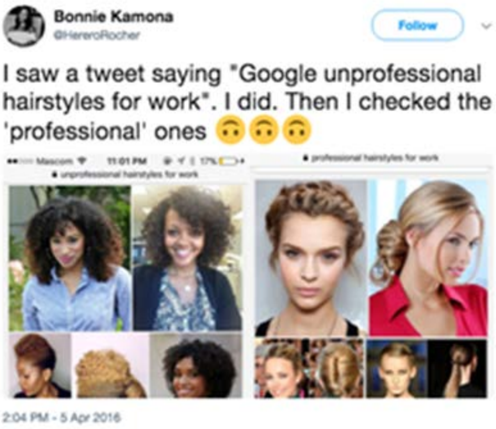
The inverse of facial recognition is represented by 'deepfake' videos. In essence, an AI model is trained on footage of a celebrity's face, and can then dynamically superimpose that face onto the body of another person, often with a verisimilitude far beyond the prior generation of faked celebrity nudes. Deepfakes gained widespread infamy in 2017 when users on the social media platform Reddit.com distributed pornographic videos with famous actors' faces superimposed on pornographic actors' bodies [58]. Forging videos of celebrities in this way raises obvious legal concerns about personal data protections, sexual and cyber-harassment, and defamation, and copyright. It is also possible that the same technology could be used to forge political speeches and soundbites, creating photorealistic videos of politicians doing and saying things that never occurred. For example, researchers at the University of Washington have developed an AI system that manipulates the video of a speaker — in their demo, Barack Obama — so that the speaker 'lipsyncs' whatever audio is provided [59].

Deepfakes aim at being non-transparent about being deepfakes, as does fake news in general. Where a society feels that fakes are degrading trust in public institutions, that society may demand transparency, just as governments might require ads that look like news articles to be labelled as ads. This could be a version of influence transparency that does not necessarily require any inspection or knowledge of the processes by which the fakes were created, although such inspection might be required in order to determine that undue influence has been exercised.

Search

In recent years a number of concerns have been expressed that the results of searches made on internet search platforms can mirror wider societal stereotypes or prejudices. This is best illustrated through examples of image searches. In June 2016 Kabir Alli posted a video of himself conducting Google Image searches [60]. He showed that when he searched for images of 'three white teenagers' this query returned results showing smiling and wholesome looking individuals. By contrast when he searched for 'three black teenagers' the query returned images of mug shots (see Figure 1a). Also in 2016, an MBA student contrasted the results she received for searches of 'unprofessional' versus 'professional' hairstyles [61] (see Figure 1b). The images suggested a difference along racial lines.





Bonnie Kamona
@HereroRocher
Follow

I saw a tweet saying "Google unprofessional hairstyles for work". I did. Then I checked the 'professional' ones 🙃🙃🙃

unprofessional hairstyles for work
professional hairstyles for work

2:04 PM - 5 Apr 2016

Figure 1: bias in image search results. a) 'three white teenagers' vs. 'three black teenagers'. b) 'professional hairstyles'

Commentators have also noticed that image search results reflect gender stereotypes. For instance a search for 'nurses' will show more women than men and a search for 'doctors' more men than women [62]. The results for a search for CEO consisted of only 11% women, at a time when 27% of CEOs in the USA were female [63].

These search algorithms are not creating content but simply ranking the content that already exists online. These kinds of stereotyped or prejudicial results are a reflection of existing societal inequities expressed in what content is posted, how often it's linked to, viewed, and tagged by users.

However, the public's reliance on search engines as a source of information has raised concerns about their influence. Some commentators have argued that search results have the capacity to alter users' perception and reinforce societal prejudices [64]. This in turn leads to questions about whether platforms such as Google therefore have a responsibility to monitor the results of their search algorithms and 'correct' them if necessary. In the section 'Conceptual Landscape I', we have referred to this as *outcome transparency.*

Algorithmic transparency may also be called for but the problem may not be with the algorithm. Plus, there is a clear trade-off here. Search platforms tend to give little information about exactly what criteria their algorithms use, at least in part because doing so would enable commercial interests to 'game' the system, making the results ranking less useful and reliable. Therefore, algorithmic transparency can be at odds with the public (and commercial) interest in producing reliable, accurate search results. This can be addressed by keeping the algorithmic inspection limited to trusted experts who are not permitted to disclose what they learn. This of course also has some risks: disclosure by accident or corruption.

There is another issue with algorithmic transparency when it comes to search engines, however. Even when the algorithms can be understood by humans, unless there are relatively clear signs of corrupt intent or wilful tampering, the processing of these algorithms are incredibly complex. How is the inspection of an algorithm going to lets us predict that fake news is going to slip through or women are going to be under-represented in searches for images of professionals? Rigorous testing of outcomes is more likely to flag a problem. The service provider could then be required to ameliorate the outcomes without requiring algorithmic transparency to the regulators or to the public. This is a case where outcome transparency, goals transparency, and influence transparency seem likely to be more effective tools than algorithmic transparency.

Personalised online content

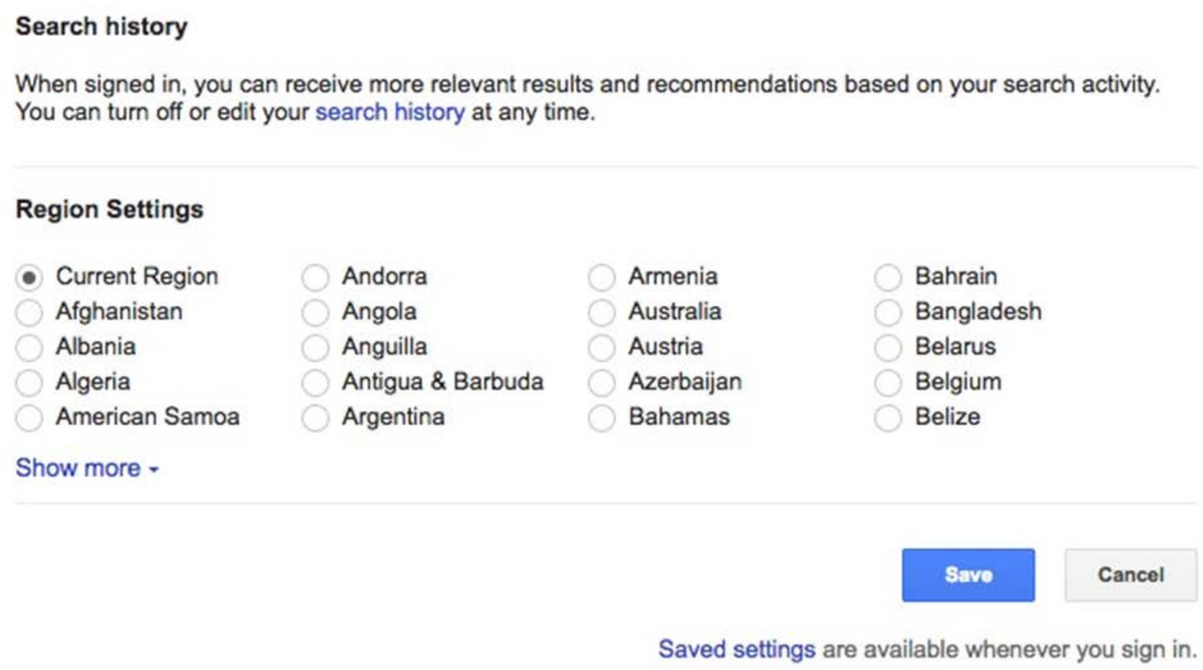
Personalisation algorithms on online platforms are designed to sift through data in order to supply users with content that is apparently most personally relevant and appealing to them. For instance, the results of a Google search may be influenced by past searches a user has made; the content and order of items in a user's personal Facebook newsfeed will be shaped by what Facebook's algorithms have calculated is of most interest to that user and Amazon recommends products based on past purchases and searches on the platform. Personalisation mechanisms therefore curate and shape much of the browsing experience [65]. This can be seen as helpful to online users as it avoids them having to sort through the vast amounts of content that are available online and instead directs them towards what they might find most useful or interesting [66]. It helps local businesses by preferring search results for services within the user's local vicinity. It also brings many advantages to internet companies as it can increase user numbers and drive up purchasing and/or advertising revenues [67]. However, concerns have been raised around the 'gatekeeping' role played by personalisation algorithms [68]. Issues include:

- **The creation of online echo chambers**. On a social network such as Facebook personalisation algorithms ensure that we are more likely to see content similar to what we have previously 'liked' or commented on. This can mean that we repeatedly see content that reaffirms our existing views and we are not exposed to anything that might challenge our own thinking [65,69]. Recent political events such as the election of Donald Trump to the US presidency have led to much debate over the role of echo chambers in modern democratic societies [70,71].

- **The results of personalisation algorithms may be inaccurate and even discriminatory.** Despite the sophisticated calculations underpinning them, the algorithms that recommend or advertise a purchase to us or present us with content we might want to see, might not in fact reflect our own interests. This can be an annoyance or distraction. More seriously, algorithms might alternatively curate content for different users in ways that can be perceived as discriminatory against particular social groups [65,72]. For instance, researchers at Carnegie Mellon University [73] ran experimental online searches with various simulated user profiles and found that significantly fewer female users than males were shown advertisements promising them help getting high paid jobs. Similarly, a researcher at Harvard [74] experimented with entering over 2000 names into the Google search platform and observing what kinds of advertisements were shown alongside the results of the search. The results indicated that searches for names associated with African-Americans were more likely to be accompanied by advertisements including the word 'arrest' (suggesting that the name in the search may be someone who has been arrested in the past) than searches for 'white sounding' names.

- **Personalisation algorithms function to collate and act on information collected about the online user.** This means that large amounts of information about individual users might be collected. Users are often unaware of the amounts of personal information being collected about them [75] or if they are, or become, aware they may feel uncomfortable about this, for instance feeling that it constitutes a breach of their privacy [76]. The impact of this perception can be seen in the emergence of options to opt out of personalisation advertisements on platforms such as Google [77] and the growth of platforms that claim not to track you [78,79].

These concerns are exacerbated by the opaque nature of most personalisation algorithms and the lack of regulation protecting the users [23]. In some cases, providers offer a weak form of algorithmic transparency by telling users at least some of what the personalisations are based on — search or purchase history, location, etc. — and may provide some control to users about what is considered in the algorithmic decisions. For example, Amazon lets users exclude purchases from consideration, and Google lets users exclude their search history from the computations (see Figure 2). These are all forms of usage transparency. In the absence of third-party verification of these explanations however concerns have been raised that incomplete transparency regarding the reasons for a personalisation outcome can be misleading [80].



Search history

When signed in, you can receive more relevant results and recommendations based on your search activity. You can turn off or edit your search history at any time.

Region Settings

● Current Region
○ Afghanistan
○ Albania
○ Algeria
○ American Samoa
○ Andorra
○ Angola
○ Anguilla
○ Antigua & Barbuda
○ Argentina
○ Armenia
○ Australia
○ Austria
○ Azerbaijan
○ Bahamas
○ Bahrain
○ Bangladesh
○ Belarus
○ Belgium
○ Belize

Saved settings are available whenever you sign in.

Figure 2: A portion of Google's 'Search Settings' page, United States version.

There are further concerns over the use of targeted advertising in connection to personalised content. This is discussed next, in the context of the Cambridge Analytica case.

Cambridge Analytica (personalisation, privacy and targeted advertising)

The 2018 'Cambridge Analytica' scandal is a high profile case that highlights particular concerns around the impact of personalisation algorithms – in particular where they relate to privacy and targeted advertising.

In 2010, Facebook launched a platform called Open Graph to third-party apps [81]. This allowed external developers to create tools that could engage with Facebook users and elicit their consent to access certain types of their personal data on the social network. This data might include the user's name, gender, location, birthday, relationship status, political and religious views, educational history and, in some instances, their private messages. The tools could also be built to allow access to the personal data of the Facebook friends of the original user.

In 2013 Aleksandr Kogan and his company, Global Science Research created an app called 'thisisyourdigitallife' [82]. The app drew on the Open Graph platform and invited users to answer a series of questions in return for receiving a psychological profile. Users were required to give permission for their personal data and their friends' personal data to be collected in order to install the app. Around 270,000 users installed it and this enabled Kogan to harvest data from 50 million Facebook profiles.

In 2014 Facebook announced changes that limited developers' access to user data [83]; these changes meant it was no longer possible for a user to give permission for third party access to their friends' data. A former manager at Facebook reported to Bloomberg that before it was discontinued, potentially hundreds of thousands of developers were making use of the third party access feature [84].

In 2015 Kogan and Global Science Research (in a breach of Facebook's policies) sold the personal data they had harvested to Cambridge Analytica, a British political consulting firm. The firm used methods based on psychometric profiling: data about individuals was collected from a variety of sources and personality profiles of them were created. Once profiled, individuals could be targeted with personalised advertisements. These would be highly tailored in terms of content and tone etc. to match the preferences of the profiled individual. Cambridge Analytica worked in support of a number of high profile campaigns, including Donald Trump's US presidential campaign and the Leave.EU campaign in the UK European Union referendum [85].

On 17 March 2018, a whistleblower exposed Cambridge Analytica's use of the harvested Facebook data [86]. Subsequent media coverage and public debate has focused on a number of issues, including two that are highly relevant to this report:

- *The ethics of the acquisition of personal data*. At the time, the use of a third party app to collect personal data from a user and his/her friends was allowed by Facebook's policies. Due to lack of usage transparency regarding these data transfers, users could prevent this occurring by changing their account privacy settings but, would be unlikely to be aware of the need to do this. Similarly, they would have been unlikely to realise that installing the app would enable such a large amount of data to be collected. As described above, users are often unaware of the extent to which online platforms and other organisations are collecting information about them for the purposes of targeted advertising. When they become aware, they can often feel that this constitutes a breach of their privacy. Awareness can also have an impact on business, with users losing trust in these organisations. For instance, the reputation of Facebook has been damaged by the revelation of their apparent willingness to allow others to access user data – something which the network has been working to address. In this case, Facebook and Cambridge Analytically arguably would have been better off being more transparent: CA might have lost some participants, but would not have been at the centre of a scandal that ultimately led to the disbanding of the company.

- *The impact and ethics of targeted advertising*. There has been much debate over the extent to which Cambridge Analytica's use of targeted advertising helped to secure victory for Donald Trump in his presidential campaign [87]. Whilst profiling techniques have been in use for a long time, their combination with algorithmically-driven personalised advertising is viewed by some as particularly troublesome. This is because, first, it allows a far greater reach across a population than other methods and might be (excessively) manipulative as individuals are unaware of how much the message of the advertisement has been tailored to their perceived preferences. Second, the success of A/B testing is evidence that people are susceptible to non-rational persuasion; placing a model to the left of a product might result in 2% more clicks than having the model on the right. AI may be able to discover even more effective non-rational triggers, which is particularly troublesome in political campaigns. Third, the high cost of this form of advertising can advantage wealthier campaigns. As a result, concerns have been raised that targeted advertising can damage the integrity of democratic institutions and Cambridge Analytica has become a symbol of the potential for political actors to make use of AI technologies in psychologically predatory ways. Fourth, the use of so called 'dark ads' that are targeted at very specific small groups of people raises concerns that they can act as secret messaging that the opposition is unaware of and therefore unable to respond to [88]. This is especially concerning in the case of negative messaging in which exaggerated, or even false, depictions of the opposing party go unchallenged because they were communicated in secret [89].

The Cambridge Analytica case is only the most prominent example of the debate over the ethics of targeted advertising online. In the business to business context, the lack of transparency of the advertising algorithms has led to concerns from companies that the automated bidding and allocation process for personalised advertisements has resulted in their products being associated with objectionable content. For instance, in 2017 a newspaper investigation revealed that advertisements for well-known brands were being placed alongside videos showing extremist content and hate speech on the YouTube platform. This risked companies being associated with the content and also a portion of the advertising revenue being passed on the video creators. Following the revelations, several companies withdraw their advertisements from the Google network. In response Google apologised and pledged to offer greater control to advertisers [90].

For these reasons, regulators may consider requiring usage transparency, along with some degree of control by users over how their data is used. As always, this will entail trade-offs possibly with the overall efficiency of the system overall (not using search histories may make the search algorithms less precise; not using automobile data might reduce the effectiveness of safety algorithms) as well as the system's performance for the individuals.

Algorithm based decision-making in the US criminal justice system

In the early 2000s the US criminal justice system began using risk assessments to assist decision-making [91, 92]. These assessments are based on algorithmic calculations to predict, for instance, how likely an individual is to re-offend or fail to attend court for sentencing. They are drawn on by the courts to help determine whether an individual should be granted bail or how long their sentence should be – with 'low risk' offenders given shorter sentences and perhaps even kept out of jail entirely. Risk assessments are now used across a wide number of states at all stages of the legal process [91]. Advocates suggest that they provide an objective measure of offender risk that overcomes potential human bias and that they can help to reduce prison overcrowding [93]. Risk assessment scores are usually made available to the defendant's legal team but the criteria through which the scores are generated are typically regarded as proprietary to the companies that develop them and are not released.

In 2016 the investigative journalism site ProPublica [93]. published a report which suggested that risk assessment algorithms might be both inaccurate and racially biased. Their journalists obtained the risk scores of 7000 people arrested in Broward County, Florida [94]. These assessments had been generated by a for-profit company called Northpointe using their algorithm, known as COMPAS. The risk scores were based on scores derived from 137 questions [93] either answered by defendants or pulled from criminal records. These questions related to factors such as personal offender history, family offender history, drug taking amongst friends and personal views on offending. Race was not one of the questions. ProPublica checked to see how many of the 7000 had been charged with new offences in the two year period since their arrest [94]. They found that:

- Only 20% of those predicted to commit a violent crime had gone on to do so.
- Of those deemed likely to re-offend, 61% went on to be arrested, when misdemeanours such as driving with an expired license were included.
- Black people were almost twice as likely to be falsely labelled as at risk of future offending than white people.
- White people were mislabelled as low risk more often than black people;
- Even when statistical tests were run to isolate the effect of race from criminal history, recidivism, age and gender, black people were still 77% more likely to be labelled as at risk of committing a future violent crime than white people and 45% more likely to be labelled as at risk of committing any kind of crime.

Northpointe countered ProPublica's analysis and criticisms [95] by stating that the algorithm was racially neutral because it had a 60% rate of accuracy for both white and black people. Since then much debate has occurred over what definitions of fairness can be applied to assess algorithms of this kind (e.g. [96]). For example, some observers (e.g. [97]) have pointed out that ProPublica claims the algorithm is unfair because it is unequally wrong (i.e. more wrong for blacks than white when making false positives) whereas Northpointe claim it is fair because it is equally right when predicting recidivism [95]. (These ideas of fairness correspond roughly to 'Equal Opportunity' algorithmic fairness [98] and 'Equal Accuracy' fairness [99]).

In 2016, a defendant who had been given a long prison sentence challenged his sentence on the grounds that he and his legal team had not been able to assess the COMPAS algorithm. However, in the Loomis vs Wisconsin case the state supreme court rejected the challenge; they reasoned that knowledge of the algorithm's output was a sufficient level of transparency and it was not necessary for defendants to know the criteria through which the scores had been calculated [100].

Algorithmic risk assessment scores continue to be used in the US to assist with decision-making in courts. In 2017 a police constabulary in the UK began to trial the approach using a machine learning tool called HART (Harm Assessment Risk Tool) [57]. In written evidence submitted to the UK government's inquiry into Algorithms Used in decision-making [101], the Head of Criminal Justice at Durham Constabulary reported that it was too early to make conclusions about the accuracy of HART, but research into it is being conducted in order to support evidence based good practice, and that the results of this research would be made available.

As will be discussed later in this section, the case raises a number of crucial points for discussion regarding the apparent lack of transparency of the COMPAS algorithm and the lack of opportunity to challenge decisions made using it.

Commentary: understanding unfairness in algorithmic systems

The case studies provide a useful means to understand the social effects of algorithmic systems for decision-making. In particular, they demonstrate the need for clear mechanisms of accountability due to their potential to bring about consequences that are detrimental on a number of levels:

- *detrimental to the individual:* individual citizens might become the recipients of inaccurate decisions (e.g., facial recognition software) or be treated more harshly in comparison to others. Where this relates to decisions over, for instance, prison sentences, this can have very serious consequences. Individuals might also receive false/misleading/skewed information e.g. as a result of online searches and this can alter their perceptions or behaviours, perhaps including their voting behaviours. The collection and collation of information necessitated by some algorithmic processes might also be considered a breach of privacy.

- *detrimental to groups:* where algorithmic processes appear to produce different results for different (demographic) groups, this often places some of those groups at a disadvantage. For instance, the case studies above suggest that blacks might be more vulnerable than whites to longer prison sentences, lack of access to facial recognition technologies, stereotyping in online advertisements, and stereotyped/prejudicial representations in online searches. This can have further detrimental consequences for those groups if the outcomes of those processes reinforce wider societal prejudices.

- *detrimental to society:* entire societies are disadvantaged if the outcomes of algorithmic processes cannot be relied on to be accurate and/or neutral. Incorrect decisions can have societal effects – for instance the wrongful arrest of individuals based on facial recognition technologies places a society at risk if actual offenders are overlooked, and stereotyped online content risks reinforcing prejudices. Furthermore, these outcomes may lead to loss of trust amongst the population as well as concerns that companies utilising these systems are allowed too much power.

Returning to the understanding of fairness outlined above, we can observe that these detrimental consequences might be considered unfair. The case studies illustrate that algorithmic processes can sometimes have social effects that do not promote equality and do not align with fundamental social values. As they also sometimes appear to hinder rather than uphold the equal distribution of societal resources, they might also be considered socially unjust. In particular, we can highlight the following social values as potentially undermined through the operation of algorithmic systems for decision-making:

- *Equality of opportunity/equality of outcome:* if algorithmic systems and/or their outcomes are biased, this may block equality of opportunity and/or outcome and systematically disadvantage certain social groups.
- *Equity:* it has been argued by some [93, 52, 74] that bias in algorithms can be discriminatory, where it disadvantages demographic groups with protected characteristics.
- *Freedom of choice.*
- *Justice:* where citizens feel that algorithms are biased or even discriminatory, this can compromise their feeling that they live in a just society.
- *Truth:* if algorithmic processes distort reality or present false information as fact, this undermines citizens' ability to determine what is true and to act on it accordingly.
- *Autonomy:* citizens' ability to act and make decisions may be undermined by various features of algorithmic processes. They may lack freedom to choose how decisions are made about them – e.g. by human vs automated process – or even to know how decisions affecting them are being made. Therefore, the lack of transparency and accountability in algorithmic systems can be particularly harmful to the societal value of autonomy.
- *Consent:* even in instances where individuals are required to give their consent to be subject to algorithmic processes, it may be that information is put to them in a way that means they are unlikely to be fully aware of what the agreement entails. The Cambridge Analytica case is a particular example of this since online users are arguably unlikely to expect that agreement to use an online app confers consent to access private messages or the details of their friends' accounts. In general, there is much concern [102] that terms and conditions on online platforms etc. are overlong and full of technical terms; even if users read them in full they may be hard to understand, meaning that genuine informed consent is unlikely. Furthermore, there is also concern regarding the mutability and dynamism of algorithms, and how the system of one-time consent to the service, even if regularly updated, cannot reflect consenting to data processing algorithms that are constantly tweaked. [e.g. 103]
- *Privacy:* privacy is fundamentally linked to consent. A key concern in the expansion of AI's scope is its effects on personal privacy, which is a fundamental human right [104]. Controversial uses include the integration of facial recognition technologies into public spaces and the degree to which advertisements online are personalised to individual users.
- *Trust*: if there is a risk that algorithmic processes may have detrimental and/or uneven outcomes, this undermines citizens' feelings of security and trust in the processes themselves and in the institutions that utilise them. This can particularly be heightened in instances where algorithmic processes are viewed as socially unjust. Further features that threaten these values are the absence of transparency and accountability; if citizens cannot see how decisions about them are being made and/or do not have opportunities to address incorrect decisions, they are less likely to trust and feel secure in these processes.

Each of these values is closely entwined with understandings of fairness and social justice. Therefore, where one or more of the values is undermined, it is possible that protests will arise stating that the algorithmic process connected to it is in some way unfair.

3.4.2. Sources of unfairness

Unfairness in algorithmic systems might result from a number of sources. Here we highlight four key potential sources: biased values in design, biased training data, biased data, inappropriate implementation of an algorithmic system.

The first type of bias could be described as biased values in design. An algorithm might be considered biased if it designed to favour one feature over another. In many cases this might be done deliberately and have trivial consequences: for example, at one point Google was thought to have reduced the weight it gave to blog posts, presumably to favour more vetted sources [105]. However, problems arise when algorithmic systems are applied to social contexts and decisions are made regarding which features the algorithm is told to associate with outcomes [106]. In extreme cases, developers intentionally construct the model so as to discriminate against certain groups and/or favour others, although such cases are presumably rare. A more common scenario is that human value judgements and assumptions play an unintentional role. Consider the design of software to filter job applications. At the heart of such software are the questions, 'What does a qualified candidate's profile look like?' and 'Who would we want to hire?' The principles that a developer decides to use to measure these qualities could introduce bias – however unintentionally. For instance, a decision that the software should favour candidates who are also Ivy League and Oxbridge graduates would disproportionately advantage white individuals and those from higher socio-economic groups. Similar implicit assumptions about what types of individuals seek certain professional roles or vote for certain political candidates could also lead to the kinds of controversies discussed in the case studies above. In these instance, socio-cultural assumptions are made by developers and embedded into the algorithm – a kind of 'values' bias. Even where particular contentious features are avoided, bias in design might still occur. In the COMPAS case it is known that race was not a specific feature included in the algorithm but it is likely that other features that were included acted as proxies for race – family history of incarceration, educational history etc. [93]. This could be a cause of the different results for black and white populations.

The absence of transparency can make it difficult to identify and assess the role of bias in algorithm design. If it is not known what features have been included in a system and why, it is not possible to trace how they might result in the disadvantageous treatment of some individuals or groups in comparison to others. By extension, this also makes the problem of bias in algorithm design harder to resolve. Goal transparency is also required to make these judgments.

A second type of bias can also occur during the development of an algorithmic system. This form stems from the data used to train the model. Most AI models 'learn' to classify unseen cases based on a 'training dataset'; this means there is potential for the model to learn from biased data and then reproduce these biases [53]. For example, a job application machine learning system trained on current data in many fields would learn that there is a low correlation between being a woman and having a job in senior management. It might well then replicate or amplify the problem. A version of this bias lies at the heart of the controversies over facial recognition technologies: the technologies were less accurate with non-white users because the training dataset did not include enough non-white faces for it to learn from. In turn, this oversight can be caused by human bias, with developers lacking awareness of the need for diversity.

The third problem source also relates to biased data. When an algorithmic system has been developed and is functioning, its outcomes might be problematic if the data it is working on is problematic. This is the problem source of concerns of outcomes of searches on Google and other search platforms. As described above, there have been multiple reported concerns over the ways in which online searchers appear to reflect, and arguably reinforce, traditional gender and racial stereotypes. For instance, a search for images of 'unprofessional hair' might show many images of black women with natural Afro-Caribbean hair [107]; searches for 'female football fans' show very sexualised images [108] and searches for 'three black teenagers' return images of prison mugshots in contrast to the wholesome images shown in searches for ‘three white teenagers’ [109]. These socially biased results are not (necessarily) caused by the algorithm design; instead the algorithms driving search platforms work through existing online content and are designed to prioritise features such as the popularity, frequency of sharing and metatagging of these existing images. These means that stereotyped assumptions about gender and race etc. are evident elsewhere across the web and are picked up and reflected by the functioning of the search engine algorithms. So, due to biased data these algorithmic systems reproduce and potentially amplify existing societal biases.

The final area relates to the application or implementation of the model. In this case unfairness results not from the design of the model itself but the way in which it is applied. Deepfakes provide an excellent example of application bias, in the sense that deepfake pornography could be considered another example of online misogyny. Neither the data used to create deepfakes (videos of celebrities) nor the algorithm design (common libraries like TensorFlow for image-based machine learning) are problematic; the issue arises in how particular—and in this case, rogue—actors apply the technology. In another instance the COMPAS case could be argued as the inappropriate application of a system as the algorithm was used in broader ways than originally designed for [93]. Similarly, sentencing AI might be used to definitively determine sentences, to recommend sentences, or to check a judge’s independent decision for bias; the gravity of the unfairness is not embedded in the AI but in its use and goals.

As we can see, the sources of unfairness identified can play a crucial part in reinforcing and perpetuating existing discrimination in society, affecting access to available resources and opportunities. Addressing these kinds of 'allocative and representational harms' is a main topic of initiatives such as the Algorithmic Impact Assessment proposal by AI Now, which we discuss in detail in (section 3.10.5.1).

3.4.3. Opportunities and barriers towards achieving fairness in algorithmic systems

Just as controversies over ‘unfair’ and ‘socially unjust’ algorithmic systems have been the focus of much debate in academic, policy and public discourses, attention has also been given to potential solutions. Various means to achieve fairness and social justice in algorithms have been suggested. As summarised below, they offer valuable opportunities for beneficial change but certain barriers exist against them being achieved.

How to understand fairness? As stated at the start of this section, fairness is a nuanced and inconsistent concept open to different interpretations. This can make overall conclusions and agreements about accomplishing fairness difficult to reach. Social research indicates [110] that when presented with a set of algorithms and asked to select their preferred one to be applied in a particular context, participants routinely select the one they feel to be the most fair for that context. However, selection differences occur because participants draw on different understandings of what constitutes fairness in that instance. In relation to algorithm design, Kleinberg et al. [111] observe that it can be impossible to fully accommodate different concepts of fairness simultaneously because they are competing rather than compatible. In their work they focused on probabilistic risk assignments and formalised three fundamental conditions that might need to be satisfied for the assignment to be considered ‘fair’. They found that it was not possible to satisfy all three constraints at the same time, except under highly constrained conditions which are unlikely to map onto to real world scenarios. They conclude that a way forward is to consider trade-offs that can be made between notions of fairness.

One of the domains that Kleinberg et al. [111] considered is the use of algorithmic decision making in the criminal justice system and they note the controversy surrounding the use of COMPAS in the US court system. COMPAS provides a highly useful case study to consider the complexities around what constitutes fairness. In the public discussions around the apparent bias of the algorithm both

ProPublica and Northpointe were able to cite evidence in support of their alternate positions. ProPublica argued that the COMPAS algorithm was unfair and potentially discriminatory because it carried a higher risk of false positives for blacks. However, Northpointe stated that the algorithm was fair because correct predictions of recidivism were equally accurate for whites and blacks. Each argument draws on different conceptualisations of fairness: unequally wrong vs equally right.

In her work on recidivism prediction instruments (RPIs) Alexandra Chouldechova [112] also draws on the COMPAS case and makes a distinction between the social and ethical conceptualisation of fairness and the statistical concepts underpinning the operation of an algorithm. She notes that the higher false positive and lower false negative rates for black defendants compared to white defendants was drawn on by critics of COMPAS as evidence that the algorithm was racially biased in its outcomes. However, she finds that this difference was a result of applying an RPI that satisfies predictive parity to a population in which recidivism rates differ – i.e. recorded rates of reoffending differed between blacks and whites. In essence, because recidivism rates are uneven across the groups it is not possible for the algorithm to simultaneously be both equally right and equally wrong for black and white defendants [113]. As a result, disparate impact (i.e. indirect discrimination) occurs, in which the RPI have disproportionate negative consequences on one (demographic) group. This impact is unintentional and may occur regardless of whether the RPI was designed to fulfil specific fairness criteria. The concept of disparate impact/indirect discrimination, and its links to GDPR, is discussed further in sections 3.10 and 4.1.

Similarly, different interpretations can be applied to the issue of search engine results. Search algorithms identify existing online content and typically prioritise – that is, give a higher ranking to – content that has been viewed most widely by users. By ranking this content higher, they then open it up to greater visibility because users are most likely to click on links that are at the top of their search results. So search engines have a multiplier effect as they make already highly visible content even more visible. This could be seen as an entirely fair process as all content is treated equally at the start of the process and sorted through the same criteria. Alternatively, it could be argued that the process is unfair as its outcome confers an additional advantage on the content that has been viewed most often. So in this case, interpretations on what is fair differ along the lines of equity of opportunity vs equity of outcome. These debates become particularly crucial in discussions of how to deal with online search results that appear to reinforce stereotyped or prejudicial views. In this case, goal transparency can at least help users evaluate the search results presented to them.

Mittelstadt et al. [35] state that algorithms 'are inescapably value-laden. Operational parameters are specified by developers and configured by users with desired outcomes in mind that privilege some values and interests over others'. Human values are (often unconsciously) embedded into algorithms during the process of design through the decisions of what categories and data to include and exclude. As already stated above, these values are highly subjective – what can appear ‘neutral’ or ‘rational’ to one person can seem unfair or discriminatory to another. So once again these subjectivities invoke different understandings of fairness. Given that algorithms have values embedded in them, transparency about those values can help users interpret the results, guard against biases, choose which systems to rely on, and engage in useful debate about the fairness of the outcomes.

From a practical perspective therefore, it is important to understand that different concepts and elements of fairness can conflict. Promoting equity in practice can be directly opposed to promoting equality; in many cases, both outcomes can be viewed as socially desirable by different actors. In a sense, defining the governing sense of ‘fairness’ is at the heart of many political disputes and is required when we operationalise fairness by programming a computer. Given the complexities and subjectivities involved, it can be hard to reach consensus in all but the most obviously unfair, harmful and discriminatory cases. As a first step toward addressing this issue however there needs to be transparency regarding the fact that a choice of choice of fairness measure is taking place.

Who is responsible for ensuring fairness?

Questions arise over where responsibility lies to resolve different forms of unfairness. For example, is Google culpable if their algorithm translates the word 'doctor' into the masculine form and 'secretary' into the feminine form in a language with grammatical gender, even though this results from the underlying data and not Google's algorithm?[114] Is Microsoft responsible when their AI chatbot learns the sexist and racist discourse of the internet? In Microsoft's case [115], they apologised but ultimately blamed 'trolls' who attacked their technology, an abdication of responsibility despite knowing that such attacks were possible and even highly likely. Should we expect governments to legislate for such matters or should we expect industries to pursue best practices? Could there be more scope for standards-setters in the industry to be more proactive in establishing best practices, rather than apologising after the case while ultimately deflecting accountability? Or could users be assumed to have enough awareness of these processes and their potential outcomes that they can exercise their own judgement, critical thinking and sense of responsible behaviour?

Multiple sources of unfairness, multiple solutions: Because unfairness in algorithmic systems has the potential to arise from a number of sources, there are multiple potential solutions to address it. This multiplicity of sources and solutions creates a challenge for ethicists and regulators, because there is no 'one size fits all' remedy and instead each type of unfairness invites different regulatory responses from industry, academia, and policymakers.

Solutions proposed to address biased design values in the development of algorithms often centre on calls for greater transparency in the development process and diversity amongst developers [116]. Controversies around biased training data have had an active response, typically focusing on efforts to ensure more inclusive datasets [117]. Some academic researchers have begun to take up the challenge to actively create training datasets that are more inclusive of different demographic groups, especially for automated language and image-related tasks [50].

Where the source of unfairness (potentially) lies in the data algorithms work on, several commentators have suggested the possibility of adjusting algorithms to accommodate for known bias. For instance, search algorithms could be tweaked to avoid ranking online content in a way that reflects gender or racial stereotypes etc. [118] or sentencing algorithms could be adjusted to ensure different subpopulations are treated evenly, even if their baseline characteristics are skewed. Finally, addressing unfairness resulting from the application of algorithmic systems is often discussed in terms of *ex post facto* laws, such as the use of anti-discrimination laws to seek redress against cases of systematic unfairness against individuals or groups with protected characteristics [119] or the expansion of the General Data Protection Regulation where it provides a 'right to explanation' of how data is processed [120]. Legislation could address algorithms in general, as an 'Internet Bill of Rights', adapting and re-affirming rights in the context of wide-scale algorithmic decision-making; this should accommodate legitimate disagreements about the various forms of fairness to bring to bear in any particular case.

Another approach, which doesn't preclude the previous one, is addressing specific algorithmic practices in legislation, as the European Union did with algorithms in High Frequency Trading [121]. Regulating the opacity of algorithms, which is mostly established through confidentiality agreements, establishing public agencies of oversight and even pushing for sectorial self-regulation (as seen in the advertising industry). A good example is New York's Bill on Algorithmic Accountability [122]. The current terms of the proposal seek to establish an agency responsible for fairness, accountability and transparency of algorithms that are used by public authorities. Citizens may solicit action from the agency in order to seek explanation and eventually contest algorithmically driven decisions by those authorities. The agency would also be responsible for policing discriminatory practices within algorithmic decision systems and providing information on how an algorithm functions and impacts the city. We return to this proposal in section 3.10.5, which discusses the proposal for Algorithmic Impact Assessment in more detail.

These different suggested solutions inevitably raise debate and are not necessarily simple to achieve in technical, practical or procedural terms. While datasets could theoretically be audited or self-regulated to combat bias resulting from training data, bias resulting from design values is generally more subjective and difficult to regulate. In many cases, transparency is neither 'necessary nor sufficient' to ascertain whether the values and intent behind an algorithmic system are biased, given the legal and social complexities associated with establishing intent. Further, it is not always possible to draw a straight line from design values to outcomes, as the COMPAS case clearly illustrates. Similarly, there is likely to be disagreement among different parties about the values at stake. One person's pragmatic efficiency might be another person's technocratic racism and many biased or discriminatory outcomes of algorithmic decision technologies began with good or neutral intentions. One person's fairness that minimises false positives may be another person's unfairness because it does not minimise false negatives. Fairness is complex.

Industry standards and the risk of fairness as an escape from regulation

As a consequence of the various controversies arising from the application of algorithmic systems for decision-making, there have been calls for industry to take the initiative to develop standards and codes of conduct to ensure fair, ethical and socially beneficial practice. Major technology companies recognise the importance of having dedicated teams looking into these challenges. Facebook's FAIR [123] is an example of an effort directed exclusively to understanding and developing artificial intelligence. Similarly, in June 2018 Google published a set of principles of AI use [124]. These are the first steps towards establishing socially-oriented goals for the development of technology, as well as benchmarks for their expected behaviour. However, in a recent article, Ben Wagner [125] warns of the limitations of these steps by describing ethics in the technology industry as an escape from regulation. He states that companies are strongly adhering to the narrative of the value of ethics as a means to avoid regulation or to display only minimal self-regulation so as to keep just above the threshold for political or legislative intervention. Claiming adhesion to 'ethics' is a way for companies to wave the flag of positive or neutral social impacts, without entailing formal institutions that could restrict their liberties.

Given these limitations, a further positive step lies in the development of standards and guidelines by industry associations. For example, the IEEE's Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems has published a report [126] on what their committees identified as the major debates around artificial intelligence and autonomous systems. The report presents issues and candidate solutions in each debate, ranging from General Principle to Design. Codes of conduct and standards provide a first-level of rules to orient development and deployment of algorithms. The benefits are the quickness and specialisation of such rules, which rely solely on industrial and technical consensus, and provide society with formalised institutions for the regulation of the activity. This is an important first step for implementation and institutionalisation of social values of fairness. One caveat though is that the provision of concrete guidance on fairness in algorithmic systems is of course also made difficult by the complexities around the various types of fairness.

3.4.4. A *Responsible* approach to the design of Algorithms

It may be that a solution to help facilitate the development of fair algorithms is to ensure a more *responsible* research and innovation process; in particular a more inclusive process. The field of Responsible Research and Innovation (RRI) emerged from concerns surrounding the societal and ethical consequences of novel technologies [127]. The notions of responsibility and fairness are core aspects of the field, and so could be seen a potential solutions in relation to the design and use of algorithms. Central to RRI is to enable an inclusive, reflexive and accountable research and innovation process. This is for the most part achieved through the development of processes and mechanisms which ensure the involvement of relevant stakeholders throughout the *entirety* of the research and innovation life cycle [128]. In relation to the development of algorithms this would likely involve a contextualised consideration of an algorithm to determine the most relevant stakeholders. Following this determination, mechanisms such as stakeholder workshops and focus groups could be integrated into the research and innovation cycle so that stakeholders could share their views with developers in a meaningful way. Importantly, those developing the algorithm would take these perspectives and concerns into account and find ways to embed them into their ongoing development. This *responsible* procedure would help ensure that the resulting algorithm would be as *fair* as possible given the real consideration and integration of multiple stakeholder viewpoints and concerns.

The design and development of algorithms through the lens of RRI would at first glance seem to help to mitigate some key issues surrounding algorithms. In particular this is because problems like those that were raised in the earlier case studies would emerge up-stream and thus could be addressed in the ongoing development of an algorithm. Potentially problematic consequences such as those related to training data sets, and the engendering of potential bias and discrimination would likely be picked up by stakeholders. However, it is important to point out that undertaking such a *responsible* process would also have its challenges. Firstly, there would likely exist a tension between transparency and accessibility to the algorithms. Issues related to transparency would invariably emerge given the proprietary nature of algorithms- how can stakeholders feasibly be involved in the development life cycle without concerns surrounding institutional privacy? Secondly, even if in the most unlikely of circumstances institutions were to make their algorithms transparent, how could these be understood by multiple stakeholders of varying technical literacy? How should such information be presented to allow stakeholders to have meaningful discussion? Other procedural issues such as fast-paced temporality in development life cycles versus the time it would take to assess and include stakeholders viewpoints, as well as costs of including stakeholders in the process would also be additional complexities to overcome in relation to this solution.

3.4.5. Fair vs. Political and Legal

The discussion around fairness should not be set in isolation of other social systems. It is important to understand that fairness is being defined within a set of existing legal and social norms and political deliberation. Therefore, the comprehension of fairness is the simultaneous inquiry of what *should* be the appropriate outcomes of algorithmic decision-making, but also a political debate which seeks the establishment of rules of conduct, producing higher order rules which detach themselves from the fairness debate.

In this regard, the first caveat to the discussion of fairness should be its relation with institutionalised norms. Mittelstadt et al. [35] have mapped ethical challenges pertaining to algorithmic decision-making, such as concerns regarding autonomy of AI, discrimination, bias, and opacity of decisions. Many of these issues are already addressed by existing rules and laws, for example prohibiting racial discrimination and anti-competitive market practices. Fairness orients action towards just outcomes, but it has to respect existing social constraints. If a constraint is deemed unfair, it can be changed though political discussion, but fairness, as a low-level institution, cannot override higher level institutions indiscriminately.

A second caveat to fairness should be the political effects of algorithmic decision-making. Fairness shouldn't be concerned exclusively with how algorithms are transforming society, for example by reinforcing discrimination, but also how these tools can shape social understanding and views. Algorithms can affect how information circulates in society, which indirectly affects social views and perceptions. This potential shaping of debate can threaten individual liberties, for example determining how users are exposed to political propaganda on Facebook. Therefore, fairness shouldn't be concerned exclusively with material outcomes, but also how algorithms organise society in a way that our values are preserved and balanced.

In this regard, elements such as transparency and accountability are not solely tools for identifying and contesting biased decisions, but they are also mechanisms that enable people to ensure the absence of threats of unfairness to society. As Rawls [42] describes, justice encompasses an overall acceptability that existing institutions generate mutual benefit and cooperation in society. Algorithms can be understood as tools that automate and enforce decisions made and encoded by the developers. These decisions are concealed from the user, and accountability and transparency enable users to access and take part in the process of agreeing on institutions that will be enforced by algorithms. Therefore, mechanisms that promote awareness, recognition and protection of social values are an essential part of fairness, since they enable people to participate, trust and ensure the reliability of algorithmic decision-making.

3.4.6. Human vs machine bias

It is also important not to consider algorithmic (un)fairness in isolation from issues surrounding human (un)fairness: more specifically human bias versus machine bias. Many algorithmic systems employed in contemporary society were initiated, at least in part, to overcome the shortcomings of human decision-making. For instance, the COMPAS algorithm was broadly welcomed when it was introduced as it was seen as a means to address the potential for bias in the reasoning of judges [129]. Similarly, HR departments increasingly use algorithmic tools to process and filter employment applications. Advocates for these systems argue that they save time and reduce human bias, in particular they are promised to be 'free from discrimination based on race, colour, religion, gender, gender identity or expression, sexual orientation, national origin, genetics, disability, age, or other factors unrelated to legitimate business interests' [129]. As already described above, the inherent value-laden nature of algorithm development can be seen to undermine such claims to neutrality. Furthermore, evaluating whether machines improve human-decision making or are superior to it, has proven to be extremely difficult to render visible given issues with *selective labelling* of data [130]. This is, a selection bias problem, where human-decision making usually foreshadows and is highly integrated with the use of the algorithm, making it difficult to quantify or measure the impact of the machine on decision-making in relation to or compared to the human counterpart. However, an important question remains: should we evaluate algorithmic systems against an ideal of perfect fairness, or is it enough for them just to be less 'unfair' than humans?

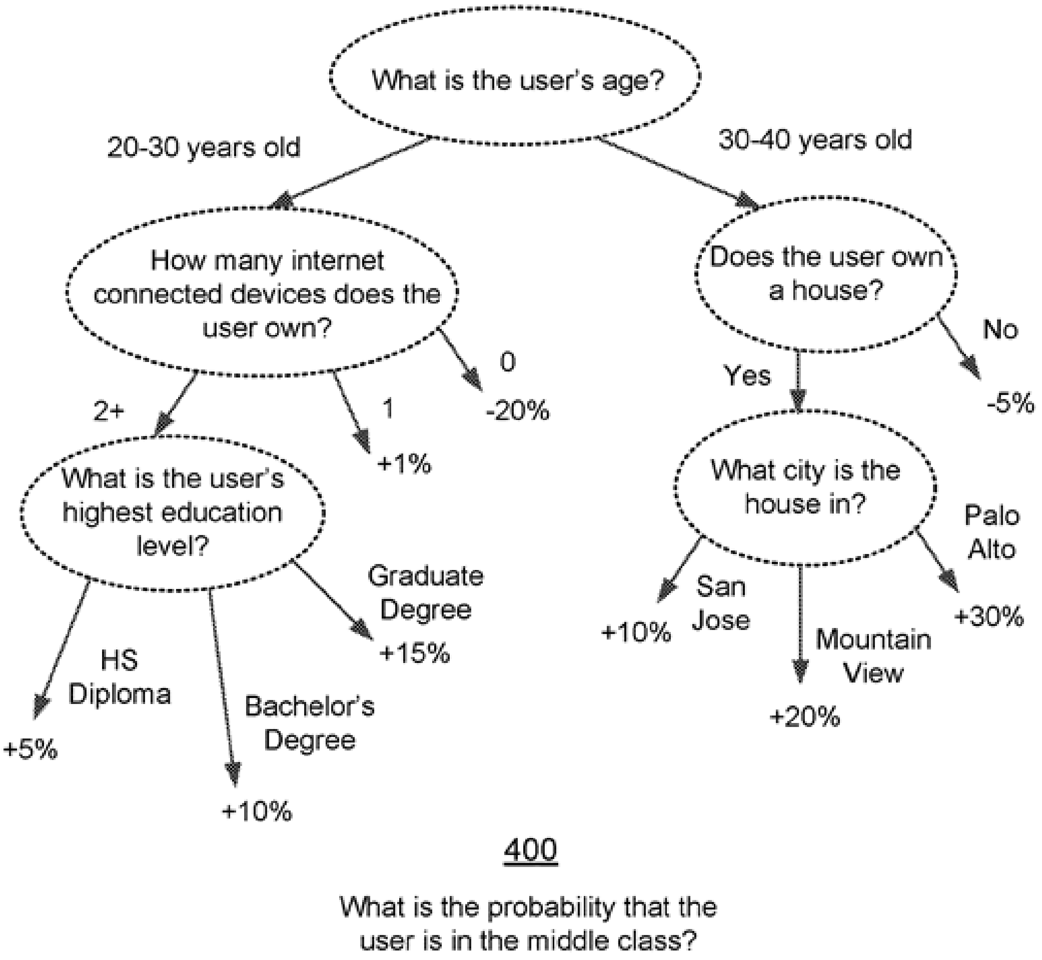
Confidentiality of industrial practices and trade secrets. As noted throughout, lack of transparency is a complicating factor in being able to assess the fairness of algorithms and also a barrier to ensuring their fairness. This particularly refers to the confidentiality of industrial practices and trade secrets. Since confidentiality is the only way of preventing other companies from copying algorithms, the protection of its method is directly related to a technology company's competitive edge. Moreover, the exposure of details about the algorithm to society could sabotage the service's own workings, since it would make it easier for users to game the algorithms to their own advantage. In legal terms, there is still no satisfactory mechanism that obliges companies to disclose their algorithms, without harming competitive strategies while ensuring public scrutiny. There is little discussion around changing IP Law in the sense of pressing for disclosure of algorithms in confidentiality agreements, though there is important literature demanding the auditability and transparency of algorithms [131].

3.5. Technical Challenges for Transparency

In order to appreciate the nature of the challenges confronting algorithmic transparency and accountability it is necessary to take into consideration the technical properties of algorithmic decision systems that can give rise to opacity [132].

We often think of algorithms as a sequence of steps arriving at an outcome, in the sense of a decision tree (see Figure 3).

Figure 3: Simple decision tree from U.S. Patent 20180032883: SOCIOECONOMIC GROUP CLASSIFICATION


What is the user's age?
20-30 years old
30-40 years old
How many internet connected devices does the user own?
Does the user own a house?
2+
1
0
-20%
+1%
Yes
No
-5%
What is the user's highest education level?
What city is the house in?
HS Diploma
+5%
Bachelor's Degree
+10%
Graduate Degree
+15%
San Jose
+10%
Mountain View
+20%
Palo Alto
+30%
400
What is the probability that the user is in the middle class?

BASED ON USER FEATURES [133]

This suggests that transparency is simple: explain the algorithm, and it will be clear how the outcome is arrived at. While strictly true, this is a naive view of the way algorithms arrive at outcomes. Some systems (e.g., expert systems [134]) are built this way, and using such a step-by-step approach may be useful. But trying to explain outcomes of most modern computing systems in this manner would not be productive. Even ignoring the intellectual property challenges of such an approach, disclosing algorithms is unlikely to be effective in providing meaningful transparency.

Key issues are complexity, interconnection of decisions, and processes that are learned from data. We now give some examples of techniques used in algorithms as well as types of algorithms, and show how these issues make explanations difficult.

3.5.1. Complexity

Modularity is often used to deal with complexity in algorithms. Different modules perform different parts of a task, and the results are combined to arrive at the final outcome. While each module may be understandable, the complexity arises as they are put together. For example, a credit scoring algorithm may include modules scoring the customer's ability to pay the total debt load, history of payment, likely profitability of a customer, etc. A consumer's record of prior payment likely goes into all of these modules, but how it affects the final credit decision may be less clear. For example, a record of prior payment may help the score based on history of payment, but since this consumes income that could be used to pay a new loan, it may decrease the score in a total debt load module.

This modularity can also be a source of error. Different modules may actually have fundamentally different meanings for the data. For example, in predictive policing some predictors might relate to the probability that a specific individual might be a criminal (each coin has a head and a trail side giving 50% chance for heads), others relate to the population probability that a certain number of people within a group will be criminals (50% of coins have two heads and 50% of coins have two tails, but if you take a specific coin it has 100% chance at a specific outcome). The latter gets included in the model as if it were the same kind of prediction as the former.

Iterative algorithms repeatedly run a sequence of steps until the algorithm converges to a stable outcome. A single pass through these steps may provide some insight into the process, but does not explain the final decision. The convergence criteria may give insight into overall goals, but may be independent of individual outcomes. For example, an advertising placement algorithm may target maximizing overall revenue through placing a package of ads, but this gives little insight into why a particular ad was shown to a particular individual.

Randomised algorithms may not run the same way every time. Often the steps to be taken are decided by a 'coin flip'. However, they are typically shown to converge to the same result each time (when given the same inputs). The randomness is introduced not to change the outcome, but to overcome problems with computational complexity. However, the fact that an algorithm may achieve the result in a different way each time makes a step-by-step explanation more challenging. It should be noted however, that software generated random numbers are typically produce through the use of pseudo-random number generators (PRNG) which generate numbers that look random but are actually part of a complex, but deterministic, number generation process. The number that is generated in this case depends on a seed to initiate the complex PRNG process. A common method for changing the seed on successive number generations with a PRNG is to derive the seed from a continuously changing input such as the internal clock. If a fixed value is used for this PRNG seed, the behaviour and results of the execution of the algorithm can be reproduced.

3.5.2. Relationship among decisions

Certain types of algorithms, particularly optimisation, are often used to solve a batch of problems simultaneously. An example would be college admissions, where there are a limited number of slots available, and the decision if an applicant should be admitted depends on how they compare to other applicants. This makes explaining any single decision difficult. Particularly challenging is when the problem involves multiple optimisation criteria, for example admitting students not just based on a computed 'score', but also based on the likelihood that they will go into particular majors.

3.5.3. Machine Learning

Machine Learning applies a very different approach to algorithm development. The basic approach is to build a model based on data. This model becomes the ``algorithm'' that is used to obtain a final result (see Figure 4).

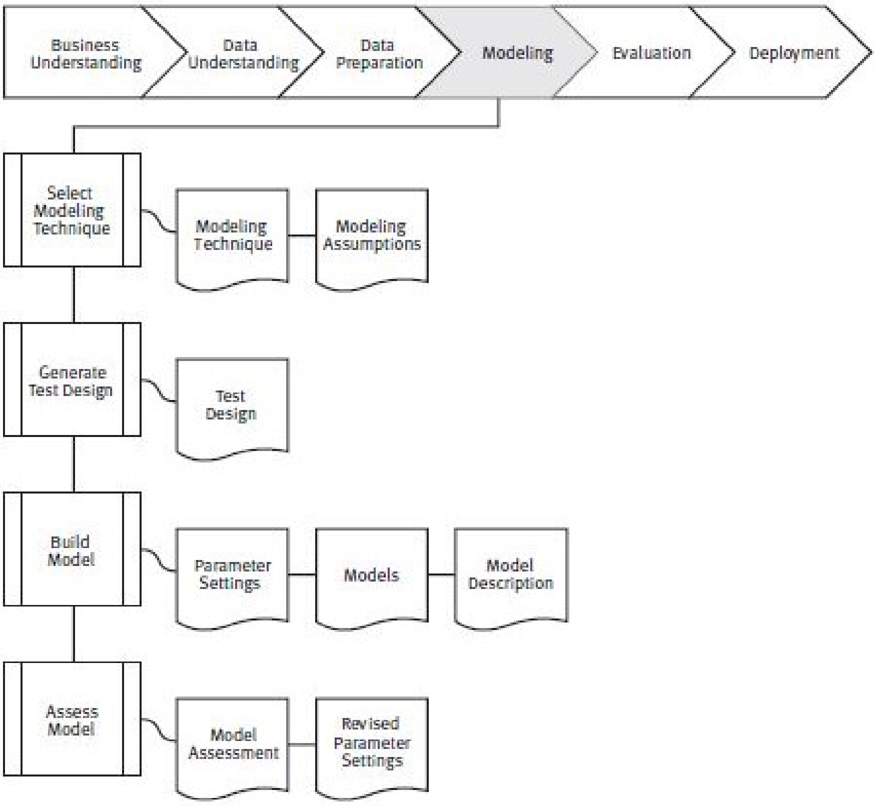


Figure 4: Machine Learning Modelling Process [135]

The machine learning algorithm itself gives little insight into the outcome; it only tells how the data is used to build the model. Transparency in the machine learning algorithm thus is of little use. What is really needed is an explanation of the model.

While the model can be viewed as an algorithm, it is typically quite complex, and often functions very differently from the way a human would make a decision. As an example, we can take decision tree learning - one of the early types of machine learning, and one of the most straightforward. Presumably we could disclose the model (decision tree), and this would provide transparency to the outcome. Unfortunately, machine learned decision trees are rarely as straightforward as that of figure 3. Figure 5 is an example of a decision tree produced using machine learning (the approach used is based on Gini coefficient for choosing splitting criteria, and minimum cost complexity pruning, see [136] for details).

Note how we keep coming back to some of the same features at different points (e.g., Uniformity of Cell Size, Bare Nuclei.) The logic behind such decisions is often unclear – and this is a relatively simple learned decision tree.

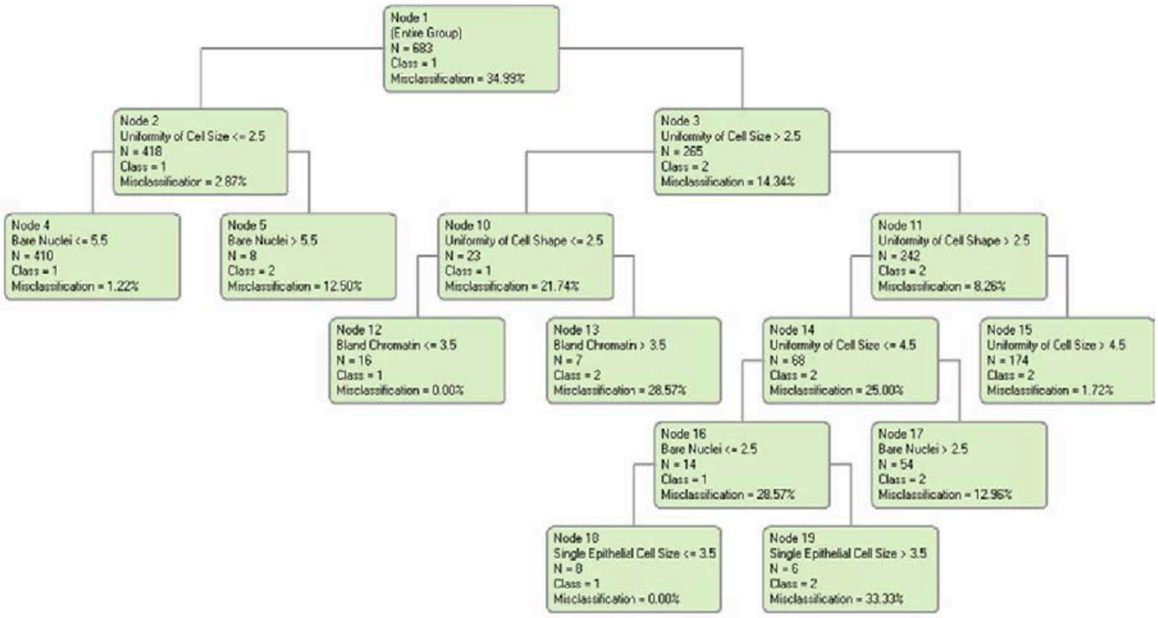

Node 1
(Entire Group)
N = 683
Class = 1
Misclassification = 34.99%
Node 2
Uniformity of Cell Size <= 2.5
N = 418
Class = 1
Misclassificatior = 2.87%
Node 3
Uniformity of Cell Size > 2.5
N = 265
Class = 2
Misclassification = 14.34%
Node 4
Bare Nuclei <= 5.5
N = 410
Class = 1
Misclassification = 1.22%
Node 5
Bare Nuclei > 5.5
N = 8
Class = 2
Misclassification = 12.50%
Node 10
Uniformity of Cell Shape <= 2.5
N = 23
Class = 1
Misclassification = 21.74%
Node 11
Uniformity of Cell Shape > 2.5
N = 242
Class = 2
Misclassification = 8.26%
Node 12
Bland Chromatin <= 3.5
N = 16
Class = 1
Misclassification = 0.00%
Node 13
Bland Chromatin > 3.5
N = 7
Class = 2
Misclassification = 28.57%
Node 14
Uniformity of Cell Size <= 4.5
N = 68
Class = 2
Misclassification = 25.00%
Node 15
Uniformity of Cell Size > 4.5
N = 174
Class = 2
Misclassification = 1.72%
Node 16
Bare Nuclei <= 2.5
N = 14
Class = 1
Misclassification = 28.57%
Node 17
Bare Nuclei > 2.5
N = 54
Class = 2
Misclassification = 12.96%
Node 18
Single Epithelial Cell Size <= 3.5
N = 8
Class = 1
Misclassification = 0.00%
Node 19
Single Epithelial Cell Size > 3.5
N = 6
Class = 2
Misclassification = 33.33%

Figure 5: Machine-generated decision tree for diagnosing Breast Cancer [132]

Decision trees are a relatively simple model. One measure of the complexity of a machine learning model is the Vapnick-Chervonkis Dimension [137]. The VC dimension of a decision tree is linear in the number of nodes in the tree [138]. Modern machine learning techniques, such as deep learning, can be much higher - the VC dimension of even a simple neural network with a binary output is quadratic in the number of parameters in the network – [139] which for a convolutional network is itself quadratic in the number of nodes in the network. As a result, machine learning models are often opaque even to their developers, and releasing the model is unlikely to provide significant transparency.

In addition to the complexity and lack of understandability of a model, there are other reasons why simply releasing a model (or the learning algorithm and the data) are not feasible solutions to transparency.

Data Privacy could be compromised. The data used to train a model is typically similar to that used by the model -- and in cases where this is data about individuals, the training data may be protected. While this would preclude releasing the data itself (and disclosing the learning algorithm without the data provides little transparency into actual decisions), it also poses problems with releasing the model. It may be possible to 'reverse engineer' a model to determine the data used to construct it, thus violating privacy [140]. While there are methods to prevent this (e.g., differential privacy [141]), using such privacy protection methods can have disadvantages, both in terms of accuracy (although the impact on accuracy is debatable [142]), and in complexity of system development.

Continuous learning also poses a challenge. In many applications, machine learning models are frequently or even continuously updated, to capture and incorporate new data and changing trends [143]. Such continuous learning models are increasingly becoming the norm in the industry, with machine learning pipelines allowing to the deployment of different models many times a day. While conceivably the model used for any particular result could be captured at that time, each model would likely require a new explanation. Given the difficulty of explaining a model, and the likely requirement for expert human involvement (at least with current technology), this further complicates issues.

3.5.4. Conclusions

Viewing transparency as 'explaining the steps of the algorithm' is unlikely to lead to an informative outcome. On the one hand, we could end up with a description that only captures the general process used to make a decision: An example of such a description, from a patent, is 'The machine learning module 250 uses machine learning techniques to train one or more models' and 'Based on the input information and the trained models, the socioeconomic group classifier 260 determines a probability that the given user belongs to the socioeconomic group.' (e.g., [144].) This provides little insight into how any individual decision is made. At the other extreme would be to provide the complete set of steps taken (e.g., the complete detailed algorithm, or the machine learned model.) While this may enable reconstructing the outcome (provided the same input data), the complexity is such that even experts may be unable to understand why a particular result was obtained.

This does not mean that the situation is hopeless. In the section on Technical Solutions we discuss alternative methods to understand the outcomes of algorithms.

3.6. Technical solutions for reducing opacity

Just as there can be technical reasons for opacity of algorithmic systems, there are technical methods for reducing algorithmic opacity, or extracting explanations for the system behaviour despite a lack of transparency.

First, we divide transparency and explanation into two categories: Understanding the overall system, and understanding a particular outcome. These may require quite different approaches. For each category, we list several approaches, and briefly discuss what each does and does not provide.

A key idea to keep in mind is the goal of transparency. Is it to understand how the system works? Or how it behaves? From a regulatory viewpoint, a primary issue is likely if the outcome is fair and appropriate – behaviour is critical. The regulatory issues governing process are likely more straightforward – GDPR, for example, forbids processing of certain types of personal information except in prescribed circumstances. This simply requires determining if such information is used by the system; in cases where use is allowed, the onus could be placed on the developer to explain and ensure that such use was proper. They key challenge, then, is transparency into system behaviour, and we should evaluate methods with respect to how they support *explanation* [145].

3.6.1. Understanding the Overall System

The goal here is to obtain a general understanding of the process by which an algorithmic system makes decisions. One challenge with the approaches described below is that they are likely to be difficult or impossible without direct involvement of system developers.

Design Review / Code Review

Design and Code reviews are methods from Software Engineering, used to enhance reliability of a system being developed and ensure that it satisfies requirements [146]. Various techniques are used, such as mapping a specific requirement to the design- and code-level modules that address that requirement. This does provide opportunity for transparency, and research showing that traditional code reviews often find issues with code understandability rather than specific defects suggest the viability of code reviews for improving transparency [147, 148].

Unfortunately, design and code reviews are expensive and time-consuming, and typically operate at a level that involves proprietary information. Furthermore, as noted in the section 3.5.4 , in a system using machine learning, this provides little transparency. The review may show that the process for building the machine learning model is as expected, but provides little insight into what that model actually will do.

Input Data Analysis

Input data analysis can be used to determine if the data being used to make decisions is appropriate and consistent with legal requirements. The process is to analyse a system at either design or code level to determine all information that is provided to a system when making a decision. This can be useful for determining regulatory compliance, e.g., a system that does not have access to race, gender, etc. as input may not be capable of direct discrimination, and thus not in violation of GDPR Article 9. This provides little insight into system behaviour, but can be a useful step provided issues with proprietary information can be resolved.

Statistical Analysis of Outcomes

For addressing some concerns, overall analysis of outcomes can be useful. For example, this can be used to identify indirect discrimination: is a protected group disproportionately affected in a negative way? The challenge is that this often requires obtaining data that would otherwise not be used. For example, a machine learning model may not make use of race or gender (avoiding direct discrimination); to store this information anyway conflicts with the principle of data minimisation and places more individual data at risk, and requiring this could potentially be considered a violation of GDPR Article 11.

An alternative approach is to create test datasets (either in a protected regulatory environment, or using synthetic data) that can be used to evaluate if overall statistical outcomes suggest there are issues. For example, standard statistical evaluation techniques could be used to determine if outcomes or accuracy are different for specific subgroups of individuals, suggesting fairness problems. This is particularly useful with static models, although it may be more difficult with continuous learning systems.

One caveat is that absolute standards for what constitutes acceptable statistical outcomes may be problematic. There have been many definitions for fairness proposed, and it has been shown that it can be impossible to simultaneously satisfy multiple definitions [37]; any hard requirement on fairness may have unintended impacts. The statistical analysis approach suggested can be useful in determining if there are large-scale issues with a system needing further exploration, rather than as a specific means of providing transparency into the decision-making process and criteria.

Sensitivity Analysis

There is also the opportunity to test systems by providing carefully crafted inputs to better understand how the systems react. For example, providing multiple test cases where only one attribute has changed can provide insight into how that attribute is used in an algorithmic decision process. This is particular important for machine learning approaches, where even the developers may have little insight into how particular decisions are made [149, 150, 151].

While a useful technique, this is by no means complete. Many algorithms, including most modern machine learning approaches, can take into account higher-order interactions between attributes. Evaluating all possible multi-way interactions is prohibitive, and as a result, such testing may fail to reveal particularly interesting cases. A potential direction arises in the development of adversarial manipulation techniques [152, 153, 154]; these can identify minimal changes that result in a different outcome, thus identifying particularly sensitive combinations of inputs.

A second issue is that care must be taken to distinguish causation from correlation. While there is a growing research literature in making this distinction [155, 156], there are still open questions, and as such results need to be used carefully.

Algorithmic Accountability

Technical issues in algorithmic accountability are largely a question if the system behaves according to specifications. Accountability issues such as redress are really beyond the technical challenges of the algorithm; these are more a question about the actions implied by the specifications. While accountability for actions taken by algorithmic systems may need to be different than for human actions, those differences are largely governed by the particular application. As a result, this section will only look at mechanisms for ensuring that algorithmic systems satisfy specifications.

Traditional software design processes include design review, code review, and testing procedures to ensure algorithmic systems meet specifications [157]. Beyond this, formal verification techniques [158] are making significant advances. Formal verification has been demonstrated on significant software artefacts [159, 160], it is likely that these techniques will become part of standard software engineering practice [161].

A second aspect of accountability is process standards and certification, such as ISO/IC JTC 1/SC7 standards for software engineering [162], or the Capability Maturity Model Integration [163]. These discuss processes and procedures organisations should follow in systems design. Within the area of algorithmic transparency and accountability, the IEEE P7000 series of standards currently under development, particularly IEEE P7001 Transparency of Autonomous Systems [164], may provide good options.

3.6.2. Transparency of Individual Outcomes

A second type of transparency is understanding a particular outcome. Here understanding how a system works is likely of little value, and approaches providing explanation become more important.

Input data analysis

Understanding what data is used to determine an outcome can be useful in establishing confidence in the fairness of an individual outcome. Furthermore, the ability to evaluate correctness of that data can identify incorrect outcomes. GDPR Article 15 already requires that data subjects have access to the personal data being processed. While this does not of itself provide explanation of an outcome, it is important to determine if an individual outcome is based on correct or incorrect data. Combined with other explanation methods, this provides useful recourse for individuals concerned about outcomes.

There are numerous cases however where access to the data that produced an outcome might not be available. Data is often considered to be a valuable asset that organisations are reluctant to share. GDPR for instance does not compel access to non-personal data, e.g. statistical data about large population groups, that might have played an important role in a decision. Furthermore, unless efforts are put in place to ensure that data is retained, for instance for data audit purposes, it might get overwritten by new inputs. A typical example where deliberate efforts are made to retain data that would otherwise disappear are flight data recorders. The mandatory inclusion of vehicle data recorders in autonomous vehicles has for instance been suggested in order to help future accident investigators get access to input data that preceded self-driving car crashes [165, 166].

Static explanation

Systems can be designed to provide explanation of the basis of individual outcomes. This can be either a specific design criteria incorporated into the entire system, or accomplished through techniques such as sensitivity analysis.

Such systems already exist in practice, even without regulatory requirements. As an example, the Fair-Isaac Corporation FICO score, commonly used in financial credit decisions in the United States, provides reports to individuals explaining their individual credit score. These provide 'the top factors that affected your FICO Score 8 from each bureau with a detailed analysis' [167]. Further, these factors have to be remediable by the individual; 'You are a woman' is not, but 'You are too often late in making your credit card payment' is.

Design / Code Review and Statistical Analysis

Techniques such as design and code review are of little direct relevance to understanding an individual outcome. However, disclosing synopses of such reviews can be part of the process of setting out 'meaningful information about the logic involved', helping to satisfy GDPR Article 15 1(h).

Sensitivity Analysis

As with overall outcomes, sensitivity analysis can be used to determine what has led to a particular outcome. By perturbing inputs — sometimes referred to as testing counterfactuals [27] — and evaluating the change in outcomes, insight can be gained into how a particular outcome has been arrived at. The ability to start with a particular set of inputs enables a wide variety of perturbations to be tried, potentially even capturing multi-variate factors. The previously discussed techniques for sensitivity analysis to study overall outcomes may provide appropriate starting points for such analysis.

Furthermore enabling sensitivity analysis for individual outcomes provides not only greater transparency, but it gives the data subject the opportunity to determine what actions might result in a different outcome, or information that can be useful in contesting an outcome. Such 'what if' analyses can provide useful information to individuals, as well as identify fairness issues that require further investigation.

In many cases, this is a tractable approach, for example, in the U.S. Fair-Isaac already offers consumers a FICO Score Simulator that shows 'how different financial decisions — like getting a new credit card or paying down debt — may affect a FICO® score' [167].

An example of a powerful, model agnostic, explanation approach for machine learning classifiers that uses input feature perturbation-based sensitivity analysis is the LIME (Local Interpretable Model-agnostic Explanations) technique [168]. LIME derives an easily interpretable model (e.g. a linear regression) that is locally faithful to the machine learning classifier in the vicinity around the individual predictions that it is seeking to explain. This is achieved by fitting the simplified model to input-output pairs that are generated by the machine classifier for input sample instances in the vicinity of the to-be-explained prediction.

3.6.3. Reverse engineering the 'Black-Box' - putting it all together

Reverse engineering the black-box relies on varying the inputs and paying close attention to the outputs, until it becomes possible to generate a theory, or at least a story, of how the algorithm works, including how it transforms each input into an output, and what kinds of inputs it's using. Sometimes inputs can be partially observable but are not controllable; for instance, when an algorithm is being driven off public data but it's not clear exactly what aspect of that data serves as inputs into the algorithm. In general, the observability of the inputs and outputs is a limitation and challenge to the use of reverse engineering in practice. There are many algorithms that are not public facing, used behind an organisational barrier that makes them difficult to prod. In such cases, partial observability (e.g., of outputs) through FOIA, Web-scraping, or something like crowdsourcing can still lead to some interesting results [169,170, 171, 172, 173].

3.6.4. Conclusions

Meaningful transparency into *how* outcomes are reached is technically challenging given modern computing systems; regulatory requirements for such transparency may significantly limit the ability to use advanced computing techniques for regulated purposes. Meaningful transparency into the *behaviour* of computing systems is feasible, and can provide important benefits. Mechanisms for behavioural transparency may need to be designed into systems, and typically require participation of the developers or operators of systems.

Fairness, Accountability and Transparency/Explainability are some of the fastest growing research areas for algorithmic decision-making systems, and especially machine learning. Not only academic funding bodies, but also industry is increasing its investment in this domain. This has resulted in the production of an increasing number of open source libraries and tools to help developers address Fairness, Accountability and Transparency requirements [e.g. 168, 174, 175].

3.7. Collateral implications of imposing Fairness, Accountability and/or Transparency requirements

Here we provide a brief overview of some of the collateral implications (i.e. indirect/secondary effects) that may occur when Fairness, Accountability and/or Transparency (FAT) requirements are imposed on algorithmic decision-making systems. A detailed analysis of these collateral implications however is beyond the scope of the current report.

Performance tradeoffs

When applied to the development of an algorithmic system, FAT requirements become additional performance criteria that modify the goals of the system optimization. The best optimization outcome as defined when including FAT requirement might therefore score lower on the fulfilment of the non-FAT system requirements than a system that is optimized without taking the FAT requirements into consideration. This is an inherent property that occurs any time a system is optimized to satisfy multiple requirements that are not fully independent of each other resulting in the need to make tradeoffs (e.g. optimizing the thickness of a motor vehicle chassis to simultaneously maximum impact strength and fuel-efficiency requirements inherently leads to tradeoffs between them). These kinds of problems are referred to in the literature as multi-object(ive) optimisation [176, 177, 178].

Trust and Trustworthiness

Trustworthiness of algorithmic systems, and trust in algorithmic performance are both vital elements for the successful growth in applications of algorithmic systems in both the public and private sectors. Both trust and trustworthiness are likely to be enhanced through the application of FAT requirement.

Trustworthiness of algorithmic systems relates to questions of reliability (predicable behaviour in normal use conditions), robustness (ability to maintain predictability in unexpected conditions), and resilience (ability to recover reliable behaviour after disruption) [179]. Transparency and/or explainability of algorithmic systems benefits these trustworthiness factors by helping to better understand how the systems behaves beyond the discrete data points provided by training/testing/validation trials. Fairness and accountability requirements more indirectly support trustworthiness due to the increased rigor of system behaviour inspection that is needed to control for fairness and establish accountability.

Trust in the behaviour of an algorithmic system refers to the human perception of the system as being worthy of trust [180]. Judgements of trust can be based on assessments regarding factors such as: the 'reasonableness' of the algorithmic outcomes, which is facilitated by algorithmic transparency/explainability; or the perceived ethical values that the service provider built into the system, which are expressed through fairness and accountability requirement.

Enhanced agency for users

Basic prerequisites for human Agency, i.e. the capacity to make a choice for oneself, are an awareness of what is going on and a capacity to meaningfully engage with the process. Compliance with FAT requirements will generally decrease the information asymmetry between citizens and service providers [181].

Impact on cost distribution

Implementing FAT requirements during system development is likely to require additional efforts in system testing/validation and potentially FAT related standards-based certification [182], to name just a few of the potential additional development costs. At the same time however, the increase rigor in during the system development may result in improved reliability, robustness and/or resilience which might reduce maintenance costs.

3.8. High-level perspective on governance frameworks for algorithmic systems

Regarding fundamental approaches (as opposed to implementations) to governance we will briefly review dichotomies between Principles vs. Rules based approaches in the context of technology related governance.

3.8.1. Principles vs. Rules based governance

Rule-based regulation prescribes in detail how to behave: 'On Dutch highways the speed limit is 120 km/hour' [183]. Rules provide certainty: when you follow a rule, you know that you will be compliant [184].

Principle-based regulation formulates norms as guidelines; the exact implementation is left to the subject of the norm: 'Drive responsibly when it is snowing' [185]. Principles provide flexibility: enables the regulatory regime to have some durability in a rapidly changing environment; and enhance regulatory competitiveness. Other stakeholders can benefit from the improved conduct of firms as they focus more on improving substantive compliance and achieving outcomes and less on simply following procedures, box-ticking or on working out how to avoid the rule in substance whilst complying with its form: ‘creative compliance’ [185].

Principles based regulation is criticised for failing to provide certainty and predictability and for allowing firms to ‘backslide’, and get away with the minimum level of conduct possible; and thus for providing inadequate protection to consumers and others [186, 187].

Most regulatory systems contain a mixture of rules and principles. Rules may become more principle-like through the addition of qualifications and exceptions, whereas principles may become more rule-like by the addition of best-practices and requirements [188,189]. In legal theory, Cunningham [190] uses three dimensions to distinguish between Principles and Rules:

1. The temporal dimension: Rules define boundaries ex ante, i.e., before adoption and implementation, whereas a Principle is settled ex post, when compliance is being audited.

2. The conceptual dimension: Rules are specific with clearly defined boundaries indicating what is, and what is not in scope. Principles by contrast are general, universal and abstract, which can lead them to appear ‘relatively vague’ [190].

3. The functional (or discretionary) dimension: Rules are defined by the regulator leaving little room for discretionary interpretation. Principles tend to give more space for interpretation to both subjects and auditors.

In addition to the general distinguishing dimensions that apply to Rules based and Principles based regulatory systems, [183] highlights an additional four characteristics that apply at the level of single rules of principles:'

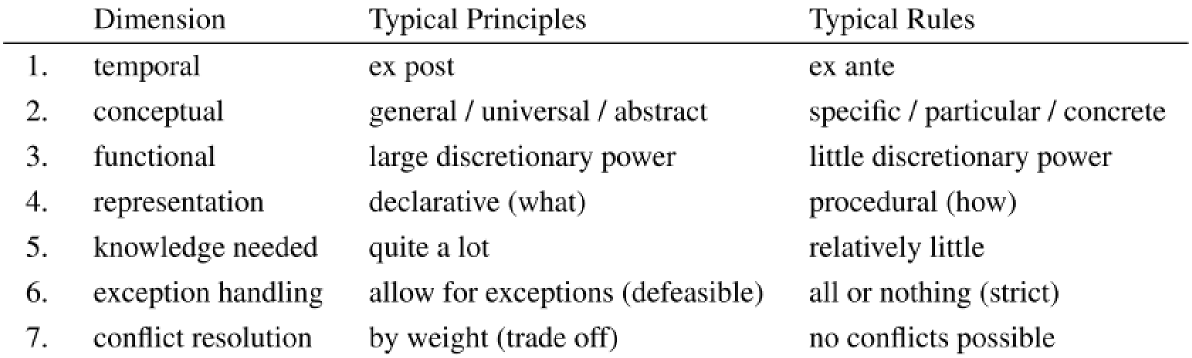
4. A declarative representation specifies what situation is required. How this should be achieved is left to the discretion of the implementer. Procedural descriptions specify how, i.e. by what actions, an objective should be achieved. Generally principles are formulated in a declarative way; typical rules are procedural.

5. What knowledge is needed to apply a regulation? Applying rules requires relatively little knowledge. Knowledge of the rule itself and the instantiation of the concepts involved, suffices. Applying principles requires more knowledge, such as knowledge of the context and of all other relevant principles.

6. How are exceptions handled? A form of reasoning may be defeasible, in the sense that exceptions may occur and overrule the original line of reasoning, or strict, in the sense that no exceptions are allowed. This can be modelled in defeasible logic [1].

7. To resolve conflicts between different exceptions we will need a kind of priority or- der or weight. In other words: for principles there is a conflict resolution mechanism; for rules no conflicts are possible.'

All seven distinguishing characteristics are summarised in Table 1.



	Dimension	Typical Principles	Typical Rules
1.	temporal	ex post	ex ante
2.	conceptual	general / universal / abstract	specific / particular / concrete
3.	functional	large discretionary power	little discretionary power
4.	representation	declarative (what)	procedural (how)
5.	knowledge needed	quite a lot	relatively little
6.	exception handling	allow for exceptions (defeasible)	all or nothing (strict)
7.	conflict resolution	by weight (trade off)	no conflicts possible

Table 1: Distinguishing characteristics of Rules and Principles by 'dimensions' [183]

For domains such as algorithmic transparency and accountability where, similar to data privacy, application conditions are rapidly and dynamically evolving, governance frameworks that are largely principles-based, such as GDPR, are likely to be more resilient to future developments than rigid rules based governance.

In order to avoid the problem of regulatory ‘backsliding’, where organisations get away with the minimum level of conduct possible, the regulatory agencies tasked to monitoring, and hence interpreting, the principles based regulation need adequate support.

In addition to being principles based, much of the existing literature/practice on governance of technology has focused on a **Risks-based approach**, i.e. emphasising methods to maximise the benefits and minimise the risks that arise from the use of the technology [191] by allocating resources in proportion to risks to society (such as health, safety or environmental risks), considering both the impacts themselves and the likelihood that they happen, in order to establish appropriate levels of control [192]. One common tool used to support risk-based approaches is an impact assessment (e.g. environmental impact assessment [193]). An example impact assessment tool that is currently being developed for algorithmic systems is presented in section 3.10.5.

3.8.2. Governance of algorithmic systems vs. systems with algorithmic components

Most algorithmic systems are not used as stand-alone systems but are part of larger integrated devices (e.g. digital personal assistant), services (e.g. credit assessment) or machines (e.g. autonomous vehicles).

Two (not mutually exclusive) ways to consider governance frameworks for these algorithms are:

1. Regulate the accountability and transparency of the algorithmic system component as considered in isolation from the rest of the system it is embedded in.

2. Regulate the accountability and transparency of the complete system, including the algorithmic and non-algorithmic component(s).

Regarding 1., there is a need to match the accountability / transparency requirement to the:

- Capabilities of the algorithmic system (e.g. different requirements for adaptive machine learning systems than for static decision rules).

- Domain where it is used (e.g. domain specific certification, i.e. algorithm X is certified to be use for product inventory tracking but not for interacting with customers).

Regarding 2., if the algorithmic system is considered as a component part of a larger system, *introduction of an algorithmic component into the system must not be allowed to break the existing accountability or transparency requirement of the governance framework that applies to the overall system* [194, 195]. For example, the inclusion of an algorithmic system for optimising parts procurement costs within a ‘just-in-time’ production process must not result in a failure to be able to trace back the origin of the parts that were used, as is required in order to be able to issue a product recall if a fault is detected in the parts from a particular supplier. This principle is already applied in various industry sectors, such as aviation where the overall system requirement for guarantees on system behaviour is imposing verification, validation, and certification challenges for adaptive flight-critical control system software [196].

A current challenge to the application of existing regulatory frameworks to systems that are significantly changed by the use of new algorithmic components is often a lack of clarity about the level of impact that the system has on citizen when a product incorporates algorithmic behaviour. For example, vacuum cleaners have traditionally been recognised to pose certain health risks through electrocution and possible dispersion of fine matter particles both of which are tested as part of the CE certification that vacuum cleaner must pass before being sold within the EEA. How does ‘upgrading’ a vacuum cleaner product line through the inclusion of autonomous navigation, camera image processing and communication algorithms, i.e. making it into a robot vacuum, change the potential for negative impact of the product? Additional tests that need to be applied to the robotic system might include issues of cybersecurity and potential for violations of privacy associated with the communications features and sensory capabilities of the device. Does the on-board algorithmic image processing extract only non-privacy sensitive information? Which information is sent over the internet? Which information is potentially accessible if the device gets hacked? [197]

The regulatory challenges introduced by the use of algorithmic decision-making as part of products or services exhibit similarities to the impact of moving to online digital service provision. In order to maintain the principle of 'online and offline equivalence' for legal and moral rights and obligations, a need to update existing laws, regulations, and international agreements to assume their application in the digital environment has been observed [198]. Similar updates may be necessary in order to maintain a level playing field where services that do, or do not, incorporate algorithmic decisions are held equally accountable for outcomes that discriminate, curtail freedom, undermine consumers' legal and moral rights etc. [198].

When considering whether algorithms should be considered separately as a single regulatory category or instead as a kind of helper technology that should be regulated as component of other kinds of technologies, Tutt [32] argues for a separate unified regulatory category on the basis that Machine Learning algorithms pose systematic complex challenges that transcend the particular technology with which they are associated. The same underlying Machine Learning algorithm could be deployed to drive a car and fly an airplane, as in the case of IBM's Watson could be used to yield expert guidance in fields ranging from medicine [199] to finance [200].

At the very least, there would need to be strong coordination between agencies when regulating algorithms in order to ensure that lessons learned in developing regulatory solutions for one set of algorithms are readily available to other agencies developing solutions to identical or highly similar algorithms. Coordination would also be necessary to provide consistency so that there is clear context based reasoning to support when the same algorithm is regulated two different ways depending on the application it is deployed in.

3.9. Governance framework options

From an institutional perspective, the governance options can be located on five stages of a continuum ranging from market mechanisms at the one end, via self-organisation by single companies, collective self-regulation by industry branches and co-regulation between state authorities and industry to command and control regulation by state authorities at the other [201, 202, 203]. Table 2 summarises these governance options and their application regarding risks associated with algorithms that mediate access to information online (e.g. search engines, news-feeds) [191].

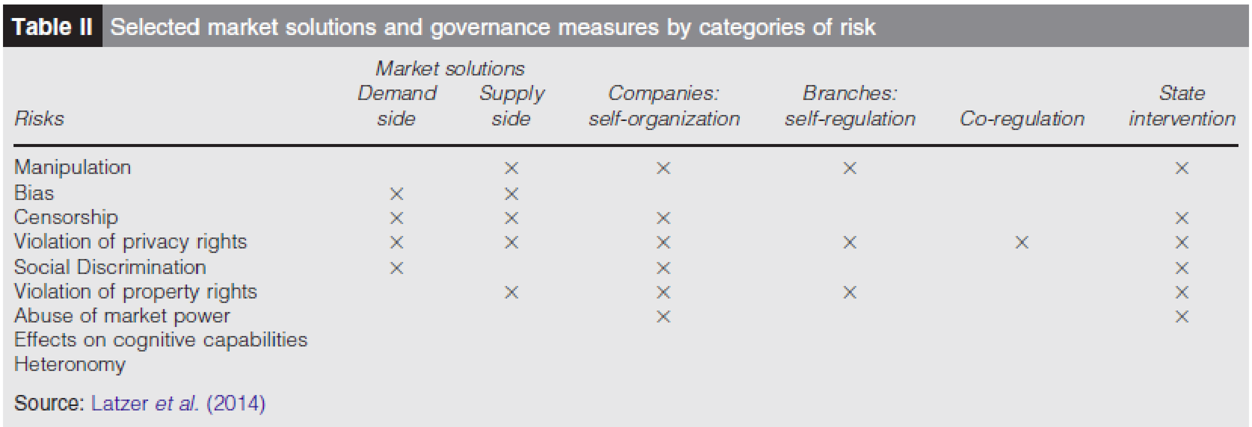


Table II Selected market solutions and governance measures by categories of risk

	Market solutions					
Risks	*Demand side*	*Supply side*	*Companies: self-organization*	*Branches: self-regulation*	*Co-regulation*	*State intervention*
Manipulation		×	×	×		×
Bias	×	×				
Censorship	×	×	×			×
Violation of privacy rights	×	×	×	×	×	×
Social Discrimination	×		×			×
Violation of property rights		×	×	×		×
Abuse of market power			×			×
Effects on cognitive capabilities						
Heteronomy						

Source: Latzer *et al.* (2014)

An important additional dimension to the governance landscape that is not explicitly highlighted in Table 2 is investigative journalism and associated public opinion shaping activities such as whistleblowers and civil-society activism that influence trust/reputation of industry and government, thus indirectly impacting on all other forms of governance [204, 205].

3.9.1. Demand side Market solutions

Demand side solutions refers to so-called market self-regulation through changes in consumer behaviour (citizen or public/private sector institutional clients) that threatens a sufficient loss of customers to motivate changes in provider conduct. In order for demand side market solutions to be viable it is imperative that alternative solutions (e.g. competing services) exist. Maintaining a plurality of solution providers however can be challenging in these markets due to inherent winner-takes-all dynamics of 'network-effects' [206, 207, 208, 209]. The performance of personalisation algorithms and machine learning systems improves with the number of data points available to the system, which is frequently linked to the number of current users [210, 211].

Based on current levels of digital literacy/algorithmic awareness and past failures to achieve significant demand side market solutions for data privacy and IoT cybersecurity, it seems highly unlikely that citizen consumer behaviour will provide a driving force for increased algorithmic transparency or accountability [212, 213, 214]. An important contributing factor to the lack of consumer action to match their privacy demands are the information and power asymmetries between service providers and consumers which often means that citizens are unaware of the extent of data collection and algorithmic manipulation they are exposed to [215, 216]. This prevents individuals from making the kinds of choice that would normally be expected to lead to market pressure for better solutions. Efforts to inform users through journalism, as well as peer-to-peer social networking, could potentially help with this, as it has with other social movements

Demand side pressure from business customers and/or public procurement are in a better position to help shape algorithmic products and services. Large brands like Unilever and Procter & Gamble have for instance put pressure on Google/YouTube and Facebook to fix aberrant behaviour of their advertising placement algorithms which had resulted in advertising for their brand products getting paired with toxic content (e.g. racism, sexism, terrorists hate messages) [217]. For the most part however such corporate activism tends to be reactive, responding to public embarrassment, threats to their brand reputation, or potential concerns about legal liability; business interests do not always align with those of customers or society more broadly

Demand side market solutions driven by public procurement by contrast are more likely to implement proactive approaches based on ethical or societal concerns [218]. When codified as pre-requisite requirements in order to bid for government contracts these take on the form of co-regulation. As an example, impact assessments currently being developed by the Canadian government [219] and proposed by AI Now [220] (see section 3.10.5 for detailed discussion) to the New York task force for examining automated decision systems used by the city [221] include public procurement as a key regulatory implementation mechanism. A weakness of governance based on public procurement requirements setting is that such measures are vulnerable to criticism and removal on the basis of fiscal responsibility requirements for reducing the cost of public procurement at the expense of ethical considerations.

3.9.2. Supply side Market solutions

Supply side solutions refers to product/service innovations that aim to capture market share by providing solutions or improvements to shortcomings of existing products/services. Examples of supply side solutions include: services that implement 'privacy by default' and 'privacy by design' to address concerns about data privacy [222, 223] and services designed to avoid filter bubbles and bias by integrating elements of serendipity [224, 225, 226].

In response to mounting number of news articles about algorithmic bias, discrimination and other ethical or security dilemmas, as well as increasing numbers of governmental inquiries, task forces and reports on these topics, some supply side market solutions are starting to emerge. Primarily these solutions are taking the form of algorithm auditing support services such as an 'AI fairness toolkit' [227, 228], 'Audit-AI' by Pymetrics [229], Facebook's 'Fairness Flow' [230] and 'ORCAA' [231] an Algorithms Auditing company set up by Cathy O'Neil, acclaimed author of '*Weapons of Math Destruction: How Big Data Increases Inequality And Threatens Democracy*' [53].

Within the academic research community there has also been an increase in research related to algorithmic transparency and accountability, as evidenced by the creation of dedicated conferences e.g. FAT* (Fairness, Accountability and Transparency)[232]; AIES (Artificial Intelligence, Ethics and Society)[233]. A number of ethics toolkits [e.g. 234, 235], and Use Case based educational material [236] are also being developed.

In order to encourage further development of such supply side solutions, however, it is vital to maintain both governmental and journalistic pressure. Corporate investment in the development of transparency and accountability of algorithmic systems is still marginal compared to the overall investment in the sector. Much of this investment represents a bet by these companies that government-mandated transparency/accountability regulations will happen in the near future.

3.9.3. Companies' self-organisation

Self-organisation refers to measures taken by individual companies to reduce risks by measures such as company principles and standards that reflect the public interest, internal quality assessment in relation to certain risks, and ombudsman schemes to deal with complaints [124, 125, 237, 238]. Self-organisation is often part of a company's broader corporate social responsibility (CSR) strategy and serves to increase reputation or to avoid reputation loss.

Company self-organisation on issues of concern regarding algorithmic systems frequently follows a similar pattern to the recent developments of supply side market solutions, where action is a response to pressures or threats to the company reputation following revelations in the news of algorithmic misconduct, or government inquiries that were perceived as threats. In the absence of such external pressure, there are hardly any incentives for companies to proactively engage in efforts for increasing transparency or accountability of algorithmic systems, unless this can be shown to contribute to improved system performance. Public disclosure of information about the workings of algorithmic systems is especially sensitive, because such disclosure increases the danger of manipulation and imitation. This results in a 'transparency dilemma' [239, 240, 241]. Moreover, a company's reputation-sensitivity affects its willingness for engaging in self-organisation measures [201, 242]. Great attention on companies in business-to-consumer (B2C) markets, such as Amazon, might promote self-restrictions in the public interest. Little public attention on companies in business-to-business (B2B) markets, such as data brokers (e.g. Acxiom, Corelogic and Datalogix)[243], reduces the reputation sensitivity and, thus, the incentives for voluntary self-organisation.

A more recent development has been an increase in (ex-)employee led activism aimed at changing company projects that are perceived to be unethical [244]. Examples include:

- Internal protest at Google against its involvement with developing AI systems for the Pentagon (Project Maven) [245] and its work on a censored search engine for the Chinese market (Project Dragonfly) [246], both of which have included resignations and/or threats of resignations by employees as well as whistleblowing to the media to generate pressure on management. As a result of these actions Google published a code of ethics to govern its AI work [247] and cancelled plans to renew its project Maven contract with the Pentagon [248].

- Ex-Silicon Valley employees speaking up, and forming pressure groups, against the use of 'addictive design' for smartphone apps and online platforms [249, 250], resulting in projects by Google and Apple to offer versions of their phone operating systems that are less addictive [251, 252].

- Employees at Amazon are demanding that the company stop selling its face recognition technology (Rekognition) to law enforcement and cancel to provision of Amazon cloud services (AWS) to the big data firm Palantir, which does work for US intelligence agencies and law enforcement [253]. Again a primary means of putting pressure on management is through speaking to the media.

Perhaps due to internal pressure, increasing public awareness, the threat of government regulation, and/or a genuine recognition of the issues machine learning poses to social fairness, major companies, including Amazon, Facebook, Google [254], IBM [255], and Microsoft, have introduced or announced open source tools for detecting systems for bias or unfairness in systems [256]. Some have also promulgated principles for the development of AI that supports social values, although those principles can be high level and difficult to enforce [e.g. 124].

3.9.4. Branches self-regulations

Typical instruments of industry/branch self-regulation are: codes of conduct [257], organisational and technical industry standards [258], quality seals and certification bodies [259], ombudsmen and arbitration/mediation boards and ethic committees/commissions [260].

Codes of conduct, such as those established by professional organisations that include software engineering in their remit (e.g. ACM [261], IEEE [262]) provide guidance on high-level principles of behaviour for the people who are developing algorithmic systems. For software development, however, codes of conduct lack enforcement power since this branch of industry, in contrast to medicine or law, is not a regulated profession that requires a license to practice. Many practitioners therefore are not member of any professional association. Corporate codes of professional conduct may carry the possibility of sanctioning and ultimately loss of the job if they are violated, but the degree to which these codes are enforced within organisations can vary widely and is often not transparent to external monitoring. Above all however, many of the ethical principles cited in these codes of conduct are very abstract and high-level, using language such as 'value alignment', 'shared propensity', 'moral responsibility', 'judicial transparency' and 'commitment to bias mitigation' [e.g. [263]) which provide little actionable help to practitioners navigating daily ethical problems in practice [264] and are subject to interpretation [265].

Academia is responding to the identified need for software developers to have a greater understanding of ethical and social implications of their work by introducing new courses such as 'Mind of the Universe – Robots in Society: Blessing or Curse? [266] (TU Delft), 'The Ethics and Governance of Artificial Intelligence' [267] (MIT), and 'Artificial Intelligence – Philosophy, Ethics, and Impact' [268] (Stanford University).

Technical and organisational/process standards [269, 270] play an important role in the software industry to ensure system interoperability (e.g. web-standards [271]), provide quality (e.g. software testing [272] and verification/validation [273]) and security (e.g. information [274] and cyber [275] security) control, good documentation (e.g. requirements [276] and user documentation [277]), maintenance [e.g. 278], review and audits [e.g. 279], and general IT governance [280]. Areas that are currently still under development include standards that expressly address ethical considerations [258], including bias [281] and transparency [161], and specific to issues related to artificial intelligence and machine learning [282]. We will discuss these ongoing efforts in more detail in section 3.11 'Development of Industry Standards relating to algorithmic systems'.

Software product certification provides a number of potential benefits to developers. It helps to establish certainty about or confidence in the software, which may stimulate sales, especially when dealing with organisational buyers in sensitive domains, like medical devices/services [283]. Certification can also help to verify and certify legislative compliance. Moreover, it can help outsourcing partners, the outsourcers as well as the subcontractor, to convince the other party that deliverables are acceptable [284]. Nevertheless, while all software goes through debugging and testing before deployment, competitive time pressure and cost considerations typically limit rigorous standards conformance certification testing to safety/security critical systems in more heavily regulated domains such as medical and aviation applications, financial services and cybersecurity. An evaluation of practices for software certification [285] reported that in the case of consumer software, being certified to the cybersecurity standard *Common Criteria* [286], time pressure often results in software vendors shipping product releases that are later versions than the one being evaluated against the standard. The same study reported that the inherently longer development times in the aviation industry meant that concerns about the time that is required for performing the certification assessment (in this case the aviation software standard *DO-178C* [287]) was much less of a limiting factor. A number of software product certification models have been developed in order to facilitate faster and more cost effective certification that software is conformant with specifications [e.g. 284] especially in the context of 'cloud' base Software as a Service (SaaS) [e.g. 288, 289]. When it comes to ethical and societal impacts of algorithmic (semi-)autonomous decision-making systems, however there are as yet no established certification models, procedures or services. Partially this is due to a lack of existing standards on these issues to certify against. There are signs however that this will change. Cathy O'Neil's 'ORCAA' Algorithmic Auditing Company has recently started to offer 'algorithmic accuracy, bias & fairness' certification [258], but certification of algorithmic systems has not yet acquired significant mainstream support.

Article 42 of the GDPR introduced the idea of voluntary certification for ML systems to demonstrate compliance with the regulation — what this report refers to as 'compliance transparency' -- although this would primarily certify compliance with data privacy principles, not absence of unjustified algorithmic decision bias [290]. Similar certification of algorithmic systems around 'big data due process' rights has also been proposed in the US [291,292], with an emphasis on two main aspects of algorithmic systems:

1. Certification of the algorithm as a software object by (a) directly specifying either its design specifications or the process of its design, such as the expertise involved (technology-based standards) and/or (b) specifying output-related requirements that can be monitored and evaluated (performance-based standards).[290]
2. Certification of the whole person or process using the system (system controller) to make decisions, which would consider algorithms as situated in the context of their use.[290]

The importance of Internal ethics committees is increasingly being acknowledged in the large technology companies as part of a response to controversies such as the DeepMind-NHS Royal Free health data transfer [293], Microsoft's 'Tay' chatbot [294] , and the 'Facebook emotion manipulation experiment' [295], which led to them to public announce the founding of a coordinated effort to develop codes of conduct for ethical AI through the Partnership on AI in September 2016 [296] (also including Amazon, Google and IBM). To what extent these ethics committees are having an impact on the work that is being done at these companies is difficult to assess from the outside, leading to some public frustration about the lack of visible impact of these ethics committees [e.g. 297, 298] and concerns that the primary function of ethics committees may in fact be to manage public image and avoid government regulation [125]. In January 2017 the Partnership on AI invited civil rights organisations (e.g. ACLU) to join in order to establish a broader multi-stakeholder platform [299]. It remains to be seen if the Partnership on AI will be any more effective than the Online Privacy Alliance (OPA), a group that formed in the mid-1990s consisting of leading Internet firms [300], to establish branch Guidelines for Online Privacy Policies [301]. The Guidelines failed to prohibit the collection of sensitive data or protect against harmful uses of data by any means other than an 'opt-out' policy [300].

3.9.5. Co-regulation

When backed by government pressure and/or monitoring, industry self-organisation and self-regulation takes on characteristics of co-regulation. This can either (1) involve objectives that are set by the regulatory body with implementation details delegated to industry (e.g. administration of the 'Right to be Forgotten' by Google [302]), or (2) involve a bottom-up approach where industry develops and administers its own arrangements, but government provides legislative backing to enable the arrangements to be enforced [300,303, 304]. This is the case, for instance, when non-compulsory rules, such as industry standards, are used as the basis for mandatory conformity certification requirements (e.g. CE marking related certification [305, 306]). An example of transparency related co-regulation are content information labels such as PEGI games content rating [307, 308]. An examples of (failed) accountability related co-regulation is the Safe-Harbour Principles for commercial data transfers between the USA and the EU, which was invalidated when the European Court of Justice (ECJ) ruled that the company self-certification practices under Safe-Harbour had failed to provide sufficient privacy safeguards for EU citizens [309, 310]. *Due to its dependence on implementation by the private sector, co-regulation is only suited to cases where fundamental rights or major political choices are not called into question* [311].

A form of top-down co-regulation which has proven itself to be effective in a number of domains are mandatory impact assessments (e.g. environmental impact assessment [312, 313]), where the regulator sets assessment criteria which the industry has to include when reporting on their assessment of the impact that is to be expected from their activities. GDPR (article 35) includes a requirement that when a type of processing using new technologies is 'likely to result in a high risk' to the rights of data subjects, then there must be a prior data protection impact assessment (DPIA), and under some conditions, consultation with the regulator. As noted by [290], DPIAs can potentially have tremendous implications for increasing transparent and accountable design/use of Machine Learning based algorithmic systems. A proposal for mandated algorithmic impact assessment as part of the procurement process is a major component of a current proposal [220] under consideration by the New York City task force that was set up to examine automated decision systems used by the city [221] (see section 3.10.5 on Accountability Measures).

In order for co-regulation to fully address concerns of lack of credibility, transparency, accountability effectiveness and enforceability of sanctions, that are frequently associated with self-regulation, co-regulation needs to establish certain guarantees and ensure a greater degree of government involvement than would be in self-regulation. Such a co-regulatory mechanism would need the following components: (1) a more balanced constitution of co-regulatory bodies with the equal participation of different partners (government, industry, and users - possibly represented by civil society groups), (2) systems that ensure that co-regulatory bodies are accountable to the government if they act outside the scope of their competences, (3) a clear, unambiguous legal basis, (4) easily accessible arrangements regarding the operation of the co-regulatory bodies, and (5) a clear division of tasks and competences between those bodies and the government [314, 315, 316, 317, 318]. The fact that the state can impose sanctions for non-compliance with established co-regulatory rules is a major difference between co-regulation and self-regulation [317].

Even if certain co-regulatory issues require resolution, a well-structured co-regulatory model has the potential to be more effectively enforced since it reduces burdens on personnel and required expertise of the regulatory body [319].

The counter argument by critics of co-regulation, as reported by [300], is that an 'industry will not reveal insider knowledge to regulators but will instead use its informational upper hand to obtain weaker standards.' [320] Moreover, the reduction in the public's opportunity to participate in co-regulatory initiatives will lead to less creativity, not more [321]. Because collaborative discussions often take place outside of the public eye, this system could also facilitate agency 'capture,' whereby government begins to pursue industry's agenda rather than the public's agenda [320, 321, 322]. Furthermore, business representatives may not enforce the rules vigorously [323], and in the absence of such enforcement, some firms may free ride on the efforts of others [320]. Established firms could also have an unfair advantage in that they could use collaborative negotiations to establish standards that discriminate against new entrants [321]. Finally, industry representatives who participate in the co-regulatory process will be conflicted because they have a strong incentive—and even a legal obligation to their shareholders—to put bottom-line concerns ahead of the public interest [321]. Critics of co-regulation express profound scepticism that this process, which gives industry a greater voice in government regulation, will yield improved social outcomes [324].'

3.9.6. State intervention

Typical state intervention instruments are: information measures to promote people's awareness and knowledge about risks, and to support appropriate behaviour; incentives by subsidies/funding (e.g. European Fund for Strategic Investments) and taxes/fees (e.g. alcohol tax); and command-and-control regulation (e.g. GDPR). An important reason for turning to state intervention instead of fully relying on the previously mentioned market led and self-governance solutions are concerns that network and scale effects are driving massive concentration in information industries [325, 326, 327], which removes market led pressures toward self-governance.

Information measures

As previously discussed, one of the contributing reasons for a lack of demand side market pressure for accountability in algorithmic decision-making is a lack of understanding by consumers/citizens regarding the ways in which algorithmic decision-making is impacting their lives [328, 329, 330, 331, 332, 333, 334, 335, 336]. The same is true for many highly skilled non-technical professionals, e.g. judges and lawyers [337, 338, 339]. Information measures can be conceptualised as consisting of two components:

1. A general understanding (i.e. 'algorithmic literacy') of algorithmic processes and the fundamentals of data analytics and machine learning is required in order for algorithmic transparency to enable accountability [340, 341, 342]. In the context of algorithmic accountability, 'algorithmic literacy' serves to provide users of algorithmic decision-making systems (e.g. public and private service providers including judges and doctors) and subjects of algorithmic decisions (e.g. citizens, customers and patients) with the basic skills necessary to critically evaluate the decisions. Without algorithmic literacy it is unreasonable to expect citizens to be able to know how, or when, they should make use of the transparency mechanisms that the GDPR [343, 344] or other laws might confer. Any understanding of algorithmic systems in the general, however will do little to provide people with the ability to judge the merits of an algorithmic decision unless it is combined with some form of public disclosure about the type and properties of the algorithms, data, goals, etc. associated with a specific decision.

2. Specific information regarding a particular application of algorithmic decision-making is required in order to make it possible for citizens, and professionals, to effectively apply their 'algorithmic literacy'. At a minimum this could take a form of algorithm/data type label notification, analogous to nutrition labels [345, 346, 347] or restaurant inspection scores [204]. Just as with nutrition labels, however, the level of information disclosure would have to be carefully calibrate to avoid problems of information overload [348] or causing vulnerabilities to manipulation by malicious actors [349]. Information that is included in a 'disclosure label' should therefore be limited to that which has the potential to either impact an individual user's decision processes, or wider public understanding of aggregate system behaviour [350]. The persistent problems with developing meaningful Privacy Policy notifications that clearly communicate data collection and handling information to citizens [351, 352, 353] stands as a warning to the challenges of operationalizing 'algorithm notifications' [354].

3. Since transparency aims not simply at making information available, but rather at making that information useful for advancing social aims, attention should be paid to the various

audiences for this information, from end-users to experts attached to regulatory bodies with regard to the level of detail, assumptions of expertise, and differing restrictions on access and re-use.

4. To make disclosed information more useful, when possible without violating privacy or trade secrets, it should be linked to sources and useful contexts. For example, a top-level user-based 'nutrition label' might link the data it presents to their underlying sources. Transparency is made more useful, reliable, and trustworthy when it does not lead to an information cul-de-sac.

Incentives by funding and taxes

Tax incentives, such as those used to boost eco-friendly technologies (e.g. electric cars, solar panels), could be used as part of an incentivising structure for promoting the use of transparency and accountability-enabling methods such as voluntary certification against transparency standards and performance auditing. This could be applied to systems of medium impact which do not qualify for investing the resources of regulatory bodies that would be required for monitoring mandatory certification. So far there does not appear to be any research into the possible advantages or disadvantages of such a scheme.

A related form of financial incentive is the aforementioned (see Demand Side Market Solutions) use of transparency requirements as part of public procurement of algorithmic services.

Strategic investment funds could be used to boost the development of new algorithmic decision-making methods that are optimised for explainability and accountable audit trails. This could be direct investment in research through Horizon2020 (and its successors), incentivising and promoting of research/innovation that solves accountability problems [355]. Some notable efforts in the US along this line are the Data Transparency Lab [356] and the DARPA Explainable Artificial Intelligence (XAI) project [357]. Part of such a strategic investment could be targeted at developing and maintaining an open source library of transparent, explainable and/or auditable algorithms with an accompanying repository of training and validation data sets. This could be done analogous to existing efforts in the research community, e.g. the International Neuroinformatics Coordinating Facility (INCF)[358]. When considering potential malicious use of AI, [359] look towards the cybersecurity community for inspiration, suggesting that the EU should financially and legally support 'red teaming' (i.e. independent groups that assume an adversarial roles to challenge organisations to improve their effectiveness) to actively probe robustness and reliability of algorithmic decision-making systems.

Another area to consider for strategic direct investment would be to provide funding for skills training, technical staff and computing infrastructure to support investigative journalism in the domain of algorithmic accountability. As was clearly highlighted by the Cambridge Analytica case [360], the controversy around the COMPAS 'recidivism algorithm' used in various US courts [94] and many more [e.g. 361, 362, 363], investigative journalism has frequently led the way in highlighting the societal implications of automated decisions (see Appendix I). To do this journalists are combining interviews, right to information requests, and investigative reporting, with computationally intensive methods (like 'black box testing,') [94, 364].

Command-and-control regulation - through legislative measures

Direct regulatory state intervention in the technology space, through legislative measures is often resisted by industry due to fears that it will limit the ability to freely explore and innovate novel technologies. Market pressures such as dominant business models and investor ‘group think’ may however result in 'innovation lock-in' which also limits free innovation and may require external pressures, such as government regulation, in order to open up new avenues of innovation [365, 366]. Prominent examples of this are in ecological technology development such as improved combustion efficiency and alternatives engine types for vehicle engines, which required government regulation (e.g. mandated catalytic converters) and subsidies to trigger the development of those improvements [367, 368]. It remains to be seen how successful the introduction of the GDPR will be for stimulating privacy as an innovation opportunity [369, 370, 371]. An important similarity between ecological considerations in engine development, privacy in online services and transparency in algorithmic decision-making (as well as cybersecurity for Internet of Things) is that despite their importance for societal wellbeing, they constitute *non-functional requirements*, i.e. requirements which are not specifically concerned with the functionality of a system [372, 373]. In the absence of state regulation, these requirements do not determine the ability of the technology to fulfil its primary design function [374].

It is tempting to look towards the GDPR (specifically article 22 often referred to as the 'right to an explanation', certain provisions of articles 13-15 'rights to ′meaningful information about the logic involved′ in automated decisions' and article 35 'data protection impact assessment') [344] and the Data Protection Authorities (DPAs) as means for enforcing algorithmic accountability. Even though the exact operationalisation of many of the clauses in the GDPR are yet to be established through legal challenges and rulings by the ECJ, various analyses by legal scholars have already pointed out that the narrow focus of the GDPR on personal data, combined with built-in restrictions (e.g. article 22 applies only to 'significant' decisions that were made with 'no meaningful human input') make it highly unlikely that the GDPR confers sufficient rights and obligations to enforce transparent and accountable algorithmic decision-making [374, 376, 378].

Beyond GDPR, the French 'loi pour une Republique numerique' (Digital Republic Act, law no. 2016-1321)[378] has drawn attention for the way it addresses algorithmic transparency and accountability [290]. The Digital Republic Act gives a right to an explanation for administrative algorithmic decisions made about individuals (so does not apply to the private sector), specifying that in the case of decisions based on algorithmic treatment (note that this includes decisions that are not fully automated but only involve algorithmic recommendations), the rules that define that treatment and its 'principal characteristics' must be communicated upon request. Further details were added by decree in March 2017 (R311-3-1-2) elaborating that the administration shall provide information about:

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

The analysis by [290] draws special attention to point 3 (above), which seems to imply the explanation must be of a particular decision (subject-based explanation) rather than a general overview of a complex model (model-based explanation). Such a focus on only the area 'local' to a specific query vastly simplifies the system that needs to be explained [379], which unlike the complexity of an entire network, might display recognisable patterns [380]. Knowing how particular factors (or 'features,' in machine learning terms) are weighted may help explain systems, but they are by no means a complete fast track to interpretability. There are at least two occasions when a court might say that weights are not useful for explaining a decision to a human user, and therefore, it is not appropriate to order disclosure. These are when the weighted inputs do not map to any real-world features the user will find intelligible and, in older or restricted systems, where retrofitting an explanation system is infeasible [290].

Several experts have proposed the use of counterfactuals as a way to assess the fairness of machine learning systems without requiring explanations. With this technique, inputs with only small differences in their data are run through the system as a way of isolating the effect of particular features. If, for example, two job applications are run through the system that differ only in the gender of the applicant, or only in that the applicant's name is changed from one typical of white Americans and the other with African-Americans, if the outcome is significantly different, then the system can be presumed to be biased without having to perform a full forensic analysis. [27, 145]

Another alternative focuses on *goal and outcome transparency*: the organisation managing a system announces what it is being optimised for, and the results of the system are monitored and possibly publicly announced; failure to meet the objectives should trigger remedial action by the organisation. This is in fact the default for AI that is not making classifications or predictions that have an immediate impact on fairness. For example, a recent project was able to predict cardiovascular risks from retinal scans, using 'deep learning,' a type of machine learning that can be even less amenable to demands for explanations [381]. If such a system were put into practice, the goal presumably would be that it diagnose those risk factors more accurately than human experts do; that goal would have to be more completely specified in terms of the levels of false positives and false negatives. If it achieves those goals, then the other forms of transparency seem unnecessary. Even if, hypothetically, the system were to turn out to assess cardiovascular risk more accurately than human experts for white patients, but dramatically less accurately than human experts for people of colour, we might still want to allow it to be used on white patients while a fix — which might not involve finding an explanation — is devised to make it useful for all people. If no such fix were found, a social decision would have to be made about whether we want to continue to use the system on white people. Note that this entire hypothetical scenario has played out with only goal and outcome transparency at play. The situation becomes more complicated in cases where particular categories of people bear a disproportionate weight of harm. For example, we might use goal transparency to ensure that autonomous vehicles are achieving social objectives such as reducing traffic fatalities and lowering the environmental impact of cars. But if those goals are being met, but the fatalities are born by a disproportionate percentage of poor people or people of colour, as an example, we are unlikely to solve the problem by prohibiting those categories of people from using AVs. Thus the regulatory disposition of the situation will be different. Nevertheless, if the inequity in the outcomes can be fixed without explanations, then this is a case where goal and outcome transparency may be judged to be sufficient. As one recent article summarised the idea: '1. AI systems ought to be required to declare what they are optimised for. 2. The optimisations of systems that significantly affect the public ought to be decided not by the companies creating those systems but by bodies representing the public's interests. 3. Optimisations always also need to support critical societal values, such as fairness.'[382]

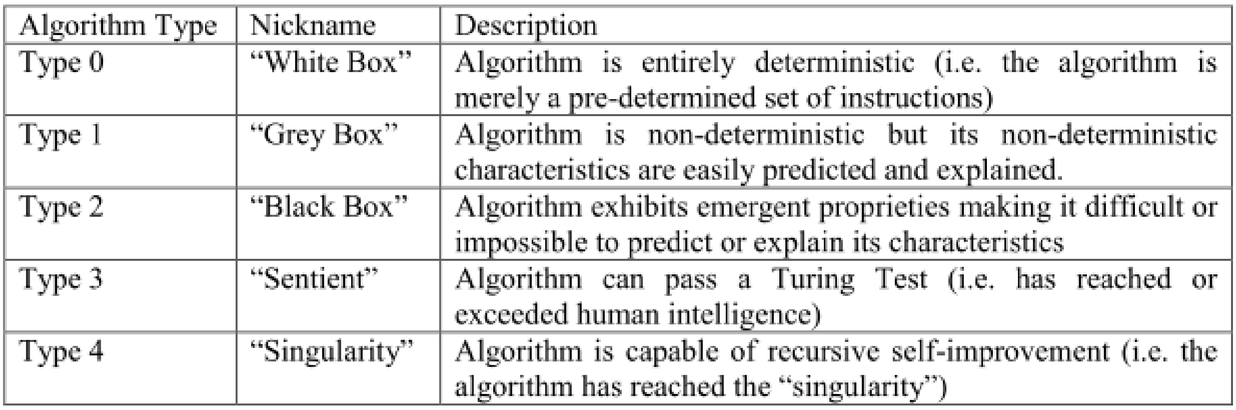
An alternative approach to algorithmic accountability that has been proposed by various legal scholars, in order to provide citizens with recourse to compensation in the absence of algorithmic transparency, is to apply No-Fault/Strict Liability to algorithmic decisions [383, 384, 385]. In order for a subject of algorithmic decisions to receive compensation under a regime of Strict Liability, the citizen would only need to show that harm had occurred (e.g. they were denied an insurance), but would not have to prove that the algorithmic decision had been faulty (hence 'no-fault liability'). Applying Strict Liability to algorithmic decision-making reverses the burden of proof, placing pressure on the organisations developing and/or using algorithmic decision-making systems to implement algorithmic transparency in order to prove that their system was not at fault (e.g. prove that the insurance was denied on valid grounds).

Command-and-control regulation - through regulatory bodies

A detailed analysis of the possible roles of a regulatory body for algorithmic systems is presented in [32]. These roles can be summarised as acting as:

1. **Standards-setting body**, possibly coordinating with Standards Setting Organisations to develop classifications of algorithmic complexity, performance-, design- , and liability-standards and best practices [386]. The algorithm complexity classification (reflecting

characteristics such as predictability and explainability) could serve to set the level of required regulatory scrutiny of decision-making algorithms. Significant scrutiny, such as a requirement for pre-deployment certification, might be reserved for the most opaque, complex, and dangerous (in term of impact on human rights and society) types of algorithmic systems—thereby leaving untouched the vast majority of algorithms with relatively deterministic and predictable outputs or lacking in significant impacts. Table 3 illustrated an example of a high-level algorithm complexity classification scheme [386].



Algorithm Type	Nickname	Description
Type 0	"White Box"	Algorithm is entirely deterministic (i.e. the algorithm is merely a pre-determined set of instructions)
Type 1	"Grey Box"	Algorithm is non-deterministic but its non-deterministic characteristics are easily predicted and explained.
Type 2	"Black Box"	Algorithm exhibits emergent proprieties making it difficult or impossible to predict or explain its characteristics
Type 3	"Sentient"	Algorithm can pass a Turing Test (i.e. has reached or exceeded human intelligence)
Type 4	"Singularity"	Algorithm is capable of recursive self-improvement (i.e. the algorithm has reached the "singularity")

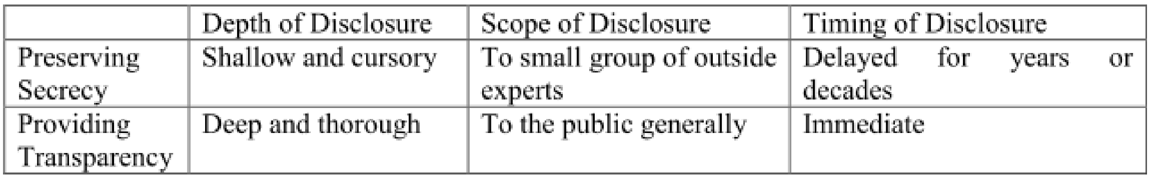
Table 3: Possible classifications of Algorithmic Complexity

Performance Standards could establish guidance for design, testing, and performance to ensure that algorithms are developed with adequate margins of safety, in accordance with its expected use, types of critical versus acceptable errors it might make, and the suggested predicted legal standard to apply to accidents involving that algorithm.

Design Standards could look to establish satisfactory measures of predictability and explainability.

Liability Standards could develop procedures for distributing responsibility for harms among coders, implementers, distributors, and end-users.

2. **Light regulator,** nudging algorithm designers by imposing regulations that are low enough cost that they 'preserve freedom of choice' and do not substantially limit the kinds of algorithms that can be developed or when or how they can be released [387]. Such 'light' regulations could involve imposing requirements of openness, disclosure, and transparency [388] that are tailored to the scrutiny classification associated with the algorithmic system (see Table 4).



	Depth of Disclosure	Scope of Disclosure	Timing of Disclosure
Preserving Secrecy	Shallow and cursory	To small group of outside experts	Delayed for years or decades
Providing Transparency	Deep and thorough	To the public generally	Immediate

Table 4: Spectrum of Disclosure [23]

3. **Hard regulation,** imposing substantive restrictions on the use of certain kinds of machine-learning algorithms, or even with sufficiently complex and mission-critical algorithms, requiring pre-market approval before algorithms can be deployed.

Pre-Market Approval. Among the most aggressive positions an agency could take would be to require that certain algorithms slated for use in certain applications receive approval from the agency before deployment. The agency could work with an applicant to develop studies that would prove to the agency's satisfaction that the algorithm meets the required performance standard.

Algorithms could also be conditionally approved subject to usage restrictions. Off-label use of an algorithm, or marketing an unapproved algorithm, could then be subject to legal sanctions. This approach may be problematic for systems that learn from their usage in the world that may be used in complex, unpredictable systems, such as autonomous vehicles.

3.10. Existing proposals for governance of algorithmic systems

The following subsections summarise a number of specific proposals related to the governance of algorithmic decision systems that have been published.

3.10.1. A right to reasonable inferences

The analysis and proposal in [389] is framed around the observation that '[c]oncerns about algorithmic accountability are often actually concerns about the way in which these technologies draw privacy invasive and non-verifiable inferences about us that we cannot predict, understand, or refute.' This is further elaborated by pointing out that '[c]ounterintuitive and unpredictable inferences can be drawn by data controllers, without individuals ever being aware [390], thus posing risks to privacy [391] and identity [392], data protection, reputation [393], and informational self-determination [394]'. This position was also summarised in [395] as 'In a big data world, what calls for scrutiny is often not the accuracy of the raw data but rather the accuracy of the inferences drawn from the data'. To further clarify this point, [395] provides as example, 'even if a bank can explain which data and variables have been used to make a decision (e.g. banking records, income, post code), the decisions turns on *inferences* drawn from these sources. Thus the actual risks posed by big data analytics and AI are the underpinning inferences that determine how we, as data subjects, are being viewed and evaluated by third parties.'

These observations are contrasted by with the focus of GDPR and proposed ePrivacy Regulation and Digital Content Directive, which aim to give 'data subjects control over how their personal data is collected and processed, but very little control over how it is evaluated' [395]. To address these accountability gaps, Tene and Polonestsky [395] propose a new right to reasonable inferences, which would be 'applicable to inferences based on non-verifiable and counterintuitive predictions which invade an individual’s privacy or damage reputation. This right would require ex-ante justification to be given by the data controller to establish whether an inference is reasonable. This disclosure would address (1) why certain data is a relevant basis to draw inferences; (2) why these inferences are relevant for the chosen processing purpose or type of automated decision; and (3) whether the data and methods used to draw the inferences are accurate and statistically reliable. An ex-post mechanism [would allow] data subjects to challenge unreasonable inferences, which can support challenges against automated decisions exercised under Art 22(3) GDPR'. One consequence of introducing this right would be that the use of Machine Learning methods, such as Deep Learning, that are currently not amenable to explanation would be ruled out for conditions where this right applies.

The proposed 'right to reasonable inferences' would focus on how data is evaluated, not just collected, apply irrespective of the identifiability of data subjects, require justification of data sources and intended inferences prior to deployment of inferential analytics at scale, and give data subjects the ability to challenge unreasonable inferences.

It is noted that such a 'right to reasonable inferences must, however, be reconciled with EU jurisprudence and counterbalanced with IP and trade secrets law as well as freedom of expression [396, 397] and Article 16 of the EU Charter of Fundamental Rights: the freedom to conduct a business.'

3.10.2. Consumer Protection authorities

In [398] and [399], algorithmic governance is considered from the perspective of consumer protection rules and the possible role of Consumer Protection authorities. Based on an analysis of the information asymmetry between consumers and service providers [400, 401, 402], Larsson [398] and deSteel [399 challenge the feasibility of asking consumers to protect themselves through consent mechanisms. Instead, [402, 403] argue for a broader application of consumer protection regulation to user agreements in order to increase accountability for operators. This in turn requires consumer protection legislation to be applied pragmatically through the responsible supervisory authorities [404]. A recommendation for consumer protection authorities is therefore to develop synergies with, in particular, data protection authorities, to provide expertise on consumer protection [398]. Transparency would likely have to include audits or control of how data-driven and targeting software operates, in order for consumer protection authorities to develop the ability to assess - in-house of perhaps through outsourced expertise - what the combination of algorithms and use of big data sources are leading to, and to discover the use of erroneous data [405]. This form of 'qualified transparency' [23] could be a way forward to keep the proprietary software and the specific design of algorithms as he business secrets they may need to be, but at the same time provide for a necessary protective mechanism against the worst case detriment to consumers [398].

3.10.3. An 'FDA' for algorithms

From an analysis of the specific challenges that come from Machine Learning type systems that are 'trained' rather than 'programmed', [32] develops a proposal for a centralised 'FDA' type regulator body for algorithms. The analogy with the US Food and Drug Administration (FDA) is motivated by the observation that explainability and predictability are not new problems. 'Technologies, such as pharmaceuticals, that operate on extremely complex systems have long confronted them. When companies begin developing drugs, their hypotheses about why they might prove effective are little better than smart guesses. Even if the drug proves effective for its intended use, it is hard to predict its side effects because the body's biochemistry is so complex. Pfizer was developing Viagra as a treatment for heart disease when it discovered that the drug is actually a far more effective treatment for erectile dysfunction [406]. Rogaine first came to market as Loniten, a drug used to treat high blood pressure before it was discovered that it could regrow hair [407, 408]. The cause of Aspirin's analgesic effects were not understood for many decades after Bayer started selling it [409]. Sometimes, once a drug is discovered, its mechanisms (including the reasons for its side effects) can be easily explained; sometimes not. But efficacy and side effects can be very difficult to predict in advance.'

The necessity for a singular new regulatory body is argued for based on three characteristics of algorithmic decision-making systems (especially those based on Machine Learning): Complexity; Opacity; and Dangerousness.

Opacity. First, the kinds of algorithms that are most concerning are by their nature opaque, with benefits and harms that are difficult to quantify without extensive expertise. That feature of the market for algorithms contrasts sharply with the market for most products where individuals are easily able to assess the benefits and safety risks posed by the product. Highly opaque and complex products benefit more from expert evaluation by a regulator than other products do.

Complexity. Second, the difficulties with assigning and tracing responsibility for harms to algorithms, and then associating that responsibility with human actors, further distinguish algorithms from other products. Algorithms could commit small but severe long-term harms or may commit grievous errors with low probability. Therefore, unlike many other products for which a combination of tort regulation and reputation will correct for accidents at an acceptable pace, the market and tort regulatory system are likely to prove too slow to respond to algorithmic harms.

Dangerousness. Third, at least in some circumstances, algorithms are likely to be capable of inflicting unusually grave harm. Whether a machine learning algorithm is responsible for keeping the power grid operational, assisting in a surgery, or driving a car, an algorithm can pose an immediate and severe threat to human health and welfare in a way many other products simply do not and cannot.

Based on these observations [32] argues that a central regulatory agency with pre-market review would be better able to contend with those problems than subject-matter agencies working independently. To the degree significant expertise is required to understand the possible dangers algorithms pose, a single central regulatory agency is more likely to be able to pool top talent together than are multiple agencies seeking to hire experts to help them make sense of the problem. A single regulator could grapple with the dangers algorithms pose holistically rather than piecemeal—effectively distinguishing between algorithms on the basis of stakeholder feedback and expert judgment. A single agency would be able to maximise the centralised expertise that can be brought to bear on the issue while offering the most agility and flexibility in responding to technological change and developing granular solutions.

3.10.4. Agency certification with tort liability

The proposal put forward by [410] constructs algorithmic accountability around an agency tasked with certifying the safety (broadly defined to include societal and discriminatory harms) of algorithmic decision-making systems in combination with a legal liability framework under which the designers, manufacturers, and sellers of agency-certified systems would be subject to limited tort liability, while uncertified systems that are offered for commercial sale or use would be subject to strict joint and several liability.[410]

The agency, staffed by specialist, would be tasked with assessing the safety of algorithmic systems while the courts, experienced in adjudicating individual disputes, would have the tasks of determining whether an algorithmic system falls within the scope of an agency-certified design and allocating responsibility when the interaction between multiple components of the system gives rise to tortious harm. This strong tort-based system would compel designers and manufacturers to internalise the costs associated with harm caused by algorithmic decisions — ensuring compensation for victims and forcing designers, programmers, and manufacturers to examine the safety of their systems.[410]

Systems that successfully complete the agency certification process would enjoy a partial regulatory compliance defence with the effect of limiting rather than precluding tort liability. Whenever a negligence suit involving the design of a certified AI system succeeds, the Agency would be required to publish a report similar to the reports that the National Transportation Safety Board prepares after aviation accidents and incidents [411].

3.10.5. Accountability Measures for Algorithmic System use by Public Authorities

Algorithmic systems are currently being used in government, reshaping how criminal justice systems work via risk assessment algorithms and predictive policing [412, 413], optimizing energy use in critical infrastructure through AI-driven resource allocation [414, 415] and changing government resource allocation and monitoring practices [412, 413]. Researchers, advocates, and policymakers are debating when and where algorithmic systems are appropriate, including whether they are appropriate at all in particularly sensitive domains [e.g. 416, 417, 418]. Questions are being raised about how to fully assess the short and long term impacts of these systems, whose interests they serve, and if they are sufficiently sophisticated to contend with complex social and historical contexts. These questions are essential, and developing strong answers has been hampered in part by a lack of information and access to the systems under deliberation. Public authorities urgently need a practical framework to assess algorithmic systems and to ensure public accountability [419].

Algorithmic Impact Assessments as a framework [419]

The Algorithmic Impact Assessment (AIA) framework is designed to support affected communities and stakeholders as they seek to assess the claims made about these systems, and to determine where – or if – their use is acceptable. It is not simply affected communities who lack the necessary information to assess how algorithmic systems are working. Governments are also struggling to assess how these systems are used, whether they are producing disparate impacts, and how to hold them accountable. Instead, impacted communities, the public at large, and governments are left to rely on what journalists, researchers, and public records requests have been able to expose [93, 420].

AIAs offer a practical accountability framework combining public authority review and public input. AIAs will not solve all of the problems that algorithmic systems might raise, but they do provide an important mechanism to inform the public and to engage policymakers and researchers in productive conversation. AIAs draw directly from impact assessment frameworks in environmental protection [421], human rights [422], privacy [423], and data protection [424] policy domains.

While AIAs resemble environmental impact assessments, data protection impact assessments, or privacy impact assessments, they differ in some very important ways. For example, data protection impact assessments (DPIAs), like those mandated under the GDPR, similarly serve to highlight the data protection risks of automated systems used to evaluate people based on their personal data [425]. If a data controller finds a system to be 'high risk,' then it must consult with its local governmental data protection authority [376]. However, DPIAs apply to both public and private organisations, are not shared with the public, and have no built-in external researcher review or other individualised challenge mechanisms. AIAs, on the other hand, are explicitly designed to engage public authorities and the people they serve on these areas of concern through the various review, public participation, and right-to-challenge elements. This allows a wide range of individuals, communities, researchers, and policymakers to participate in accountability efforts.

Pre-acquisition review

An AIA covers any algorithmic system before it is deployed, no matter how it was acquired or if it was developed internally. A pre-procurement AIA gives a public authority the opportunity to engage the public and proactively identify concerns, establish expectations, and draw on expertise and understanding from relevant stakeholders. Although framework agreements are often used as an easier alternative to complying with EU Procurement Directives, they are not an ideal procedure for acquiring algorithmic systems because they would hinder the ability of the public and government to identify and address concerns.

An AIA gives the public the authority and the opportunity to evaluate the adoption of an automated decision system before the public authority has committed to its use. This allows the public authority and the public to identify concerns that may need to be negotiated or otherwise addressed before a contract is signed. This is also when the public and elected officials can push back against deployment before potential harms can occur. In implementing AIAs, authorities should consider incorporating AIAs into the processes they already use to procure algorithmic systems or any existing pre-acquisition assessment processes the public authority already undertakes [426]. Finally, pre-procurement AIAs may also allow member states to identify relevant training and policy architecture in accordance with the European Commission's Recommendation on the professionalisation of public procurement [427].

Creating a Definition

In an AIA process, public authorities must first publish their own definition of 'automated decision system' that is both practical and appropriate for its particular context. This does not mean the public authority must go through the effort of redefining 'automated decision system' for each particular system: once they reach a working definition, they can choose to republish it in future

AIAs as long as it continues to accurately describe the systems in ways that reinforce public trust and accountability. Public authorities should also regularly revisit their definition when necessary to incorporate new types of systems, new applications of old systems, or research advances in relevant fields.

This process of defining and specifying algorithmic systems would help build the public authority's capacity for the procurement and assessment of future systems; experience with AIAs would help guide budgeting and other key milestones in the acquisition process.

Decisions about which algorithmic systems should be subject to an AIA process will be particular to each public authority's context and the interests of the communities they serve. An overly- broad definition could burden authorities with disclosing systems that are not the main sources of concern. If a public servant uses a word processor to type up her notes from a meeting where some key decisions were made, and then checks them with the program's 'automated' spell-checker, her public authority should not have to perform an AIA for that spell-checker. Alternatively, an overly-narrow definition could undermine efforts to include high profile systems like those deciding where students go to school or how housing opportunities are allocated [428, 429]. In the UK, a review of 'governmental analytical models' focused on models that are used to inform public authority decisions. The review, which went on to inform the UK Government's 'Aqua Book' on guidance for producing quality analysis in government, offers one possible method for defining automated decision system [430].

It is also essential that 'systems' are defined in terms that are broader than just their software: AIAs should address human and social factors, the histories of bias and discrimination in the context of use, and any input and training data [431, 432]. Bias in algorithmic systems can arise as much from human choices on how to design or train the system as it can from human errors in judgment when interpreting or acting on the outputs [433]. Evaluating a risk assessment tool, for instance, is not just a matter of understanding the math behind an algorithm; we must understand how judges, police officers, and other decision-makers influence its inputs and interpret its outputs [434].

In the GDPR, automated profiling is defined as 'any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.' [435]

The GDPR language is a good starting point for some authorities, but will require some shaping to match the appropriate contexts. In other contexts it may not be sufficient. Some predictive policing tools, for example, do not necessarily 'profile individuals', and instead focus on locations, using statistics to try to understand and predict crime trends across geographical areas, with the potential for disparate impact. A definition might then have to account for 'any systems, tools, or algorithms that attempt to predict crime trends and recommend the allocation of policing resources' in non-individualised terms. In general, any definition should certainly cover systems that might have a disparate impact on vulnerable communities and to pay careful attention to how broad terms, like 'automated processing,' are specified in practice. Public authorities can also learn by borrowing definitions from other domains and governments that are better tested, already have public approval, or perhaps have even withstood challenges in court.

Archive of Systems Decisions

Once a public authority has published their definition of an automated decision system, it will need to determine which systems meet this definition and will be subject to AIA requirements. Even when a public authority strikes the right balance with its definition of an automated decision system, subsequent systems decisions can be subjective or influenced by other considerations, such as intentional avoidance of performing AIAs. Therefore, public authorities should keep a public archive of all systems decisions.

The public archive can provide much needed transparency of public authority decisions, so that the public is both aware of and can challenge decisions where an automated decision system is improperly excluded from AIA requirements. Such transparency mechanism are not novel within the impact assessment frameworks. For instance, in the US Environment Impact Assessment context, a proposed federal action can be excluded from a detailed environmental analysis if it is determined that the action will not 'individually or cumulatively have a significant effect on the human environment.' Each public authority develops procedures to perform this assessment, which often take the form of a detailed checklist or determination document explaining the exclusion [436]. Additionally, some federal authorities maintain a database of all decisions that resulted in a proposed actions being excluded from further environmental analysis [437].

Public authorities can determine the level of detail of the archive, but at minimum it should provide adequate documentation of a public authority's decision to exclude systems from the AIA requirements. The public authority should also publish decisions of excluded systems before the system is deployed, so there is a meaningful opportunity for public scrutiny.

Public Authority Assessment of Systems

AIAs increase the internal capacity of public authorities to better understand and explicate potential impacts before systems are implemented [438, 439, 440]. Public authorities must be experts on their own algorithmic systems if they are to ensure public trust. This is why public authorities' AIAs must include an evaluation of how a system might impact different communities and a plan for how authorities will address any issues, should they arise.

Ideally, public authorities should pre-identify issues and potential harms that will be evaluated in the self-assessment. For example, in 2014, Former Attorney General Eric Holder urged the Sentencing Commission to 'study the use of a data-driven analysis in front-end sentencing - and to issue policy recommendations based on this careful, independent analysis.'[441] By standardising the process, authorities can ensure the evaluation is comprehensive and comparable. The evaluation should be detailed so that outside researchers and experts can adequately scrutinise the system and its potential impact, and provide a non-technical summary for the general public. This dual explanation is used in other types of impact assessment frameworks and encourages robust public engagement [442].

In their self-assessments, public authorities should identify potential impacts on the public and then proactively engage affected communities to ensure that a system meets a given community's goals. The assessment should articulate why, in light of these goals, the system will have a net positive impact on those communities [443]. Fulfilling this requirement of the AIA process would require a public authority to engage those communities early on, even before the formal notice and comment process.

Authorities could also use the AIA as an opportunity to lay out any other procedures that will help secure public trust in such systems. If appropriate, the public authority might want to identify how individuals can appeal decisions involving algorithmic systems, to make clear what appeals processes might cover a given system's decision, or to share its mitigation strategy should the system behave in an unexpected and harmful way [444, 445]. If a harm, an undesirable outcome, or an error is identified, the public authority should explain how it intends to correct or remedy the issue.

This self-assessment process is also an opportunity for public authorities to develop expertise when commissioning and purchasing algorithmic systems, and for vendors to foster public trust in their systems. authorities will be better able to assess the risks and benefits associated with different types of systems, and work with vendors and researchers to conduct and share relevant testing and research on their automated decision system, including but not limited to testing for any potential biases that could adversely impact an individual or group. Indeed, researchers are already developing resources and materials that authorities can use to ask appropriate questions of their own systems [446]. As noted above, if some vendors raise trade secrecy or confidentiality concerns, those can be addressed in the AIA, but responsibility for accountability ultimately falls upon the public authority.

The benefits of self assessments to public authorities go beyond algorithmic accountability: it encourages authorities to better manage their own technical systems and become leaders in the responsible integration of increasingly complex computational systems in governance.

Benefits to Vendors

AIAs would also benefit vendors that prioritise fairness, accountability, and transparency in their offerings. Companies that are best equipped to help authorities and researchers study their systems would have a competitive advantage over others. Cooperation would also help improve public trust, especially at a time when scepticism of the societal benefits of tech companies is on the rise [447]. These new incentives can encourage a race to the top of the accountability spectrum among vendors.

Benefits to public records request processes

Increasing public authority expertise through AIAs will also help promote transparency and accountability in public records requests. Today, when public authorities receive open records requests for information about algorithmic systems, there is often a mismatch between how the outside requestor thinks authorities use and classify these technologies and the reality [448]. As a result, requests may take a scattershot approach, cramming overly broad technical terms into numerous requests in the hopes that one or more hit the mark. This can make it difficult for records officers responding in good faith to understand the requests, let alone provide the answers the public needs.

Even open records experts who are willing to reasonably narrow their requests may be unable to do so because of the lack of any 'roadmap' showing which systems a given public authority is planning, procuring, or deploying. For example, in a project at the University of Maryland, faculty and students working in a media law class filed numerous general public records requests for information regarding criminal risk assessment algorithm usage in all fifty US states [449]. The responses they received varied significantly, making it difficult to aggregate data and compare usage across jurisdictions. It also revealed a lack of general knowledge about the systems among the authorities, leading to situations where the students had to explain what ‘criminal justice algorithms’ were to the public servants in charge of providing the records on their use. Accountability processes such as the AIA would help correct this mismatch on both sides of the equation.

Researchers, journalists, legal organisations, and concerned members of the public could use AIAs to reasonably target their requests to systems that were enumerated and described, saving public records staff significant time and resources. Public authority staff would also gain a better understanding of their own systems and records and could then help requestors understand which documents and public records are potentially available. This alignment would increase efficiency, lower the public authority burden of processing requests, and increase public confidence. And of course, some basic requests will be preempted by the AIA’s disclosure requirement, saving researchers and the authorities the burden of engaging in the public records request process.

Considering allocative and representational harms

An anticipated challenge for governments performing AIAs is the assessment of potential cultural and social harms. This challenge exists in other impact assessment processes because it requires the public authority to make assumptions or predictions about cultural or social factors that vary enormously within and between communities and geographic areas. This practice often results in findings only reflecting potential impacts on a dominant culture and omitting or misinterpreting the impacts on marginalised communities and individuals. For instance, in France, protests in predominantly Black and Muslim neighbourhoods were often represented as act of criminality and delinquency, rather than responses to poverty, exclusion, and abusive police power [450, 451, 452]. This prevailing viewpoint unfortunately resulted in greater police presence and questionable police practices in these neighbourhoods until France's highest court ruled that police racially based ID checks were illegal and discriminatory [453]. Avoiding these sorts of harms is a key goal of the AIA public participation process.

The existing literature on bias in algorithmic systems has tended to rely heavily on what could be called 'harms of allocation,' in which some groups are denied access to valuable resources and opportunities [454, 455]. Of course, addressing allocative harms is crucial. But authorities should also consider harms of representation – the way a system may unintentionally underscore or reinforce the subordination of some social and cultural groups. For example, researchers classify Google's photo platform's automatic labelling of images of black people as 'gorillas' as a representational harm [456], and the denial of mortgages to people who live within a particular zip code as an allocative harm [457]. Algorithmic systems used in the public sector are susceptible to both kinds of harm because they can be embedded with demographic data that serve as proxies for particular groups or reinforce past harms that can have economic or identity-based impacts.

Datasheets requirement

The technical research community studying fairness, accountability, and transparency in machine learning has begun to consider standard ways to account for data and its history, biases, and skews. One such proposal has called for data creators to produce 'datasheets for datasets.'[458] A datasheet is a semi-structured document that asks questions like 'Why was the dataset created?,' 'How was the data collected?,' or 'If it relates to people, does it unfairly advantage or disadvantage a particular social group?'. From those prompts, the data creator can publish information about data's provenance, its biases, and its potential societal impacts.

Datasheets could become a part of an AIA in one of two ways. First, if a public authority's acquisition or development of an algorithmic system requires the use of government data or the collection of new data, the AIA process could require the public authority to create a datasheet. Second, if the public authority purchases an algorithmic system from a third-party vendor, the vendor's contract with the authority could require the vendor to provide the public authority with datasheets for any data used by the vendor in the development of the system. While the datasheets alone would not meet the level of analysis needed in the public authority self-assessment, they are potentially valuable public documents that the public authority could include in their AIA.

PUBLIC PARTICIPATION

A fundamental aspect of government transparency and accountability is notice of how our rights may be affected by government agencies and actors. When algorithmic systems play a significant role in government decisions, the public should be given notice. Substantive public engagement requires access to accurate and timely information. Thus, an important component of an AIA is for each public authority to publicly disclose proposed and existing automated decision systems, including their purpose, reach, internal use policies, and potential impacts on communities or individuals. This requirement by itself would go a long way towards shedding light on which technologies are being deployed and where accountability research and community advocacy should be focused. AIA disclosures would also help governments proactively avoid political turmoil and backlash involving systems that the public may ultimately find untrustworthy or that may cause direct or indirect harm.

It also provides an opportunity for meaningful public participation. The AIA process includes the opportunity for the public to engage with the public authority regarding the content of its initial AIA disclosure. Public authorities can decide how they want to organise the public participation process: they could choose to separate each component of the AIA ('definition,' 'disclosure,' 'self-assessment,' and 'meaningful access') into separate public participation periods or release the AIA as a single document and have one overarching public participation period for that one document. There might be an advantage to agencies and the public in separating the definition of automated decision systems and the disclosure of systems before moving on to discuss internal assessments and external researcher access protocols. The initial disclosure provides a strong foundation for building public trust through appropriate levels of transparency, while subsequent requests can solicit further information or the presentation of new evidence, research, or other inputs that the agency may not have adequately considered.

Creating and Implementing Mitigation and Corrective Measures

A public authorities' self- assessment should identify potential impacts on the public and then proactively engage affected communities to ensure that a system meets a given community's goals. The assessment should articulate why, in light of these goals, the system will have a net positive impact on those communities [459].

Public authorities could also use the AIA as an opportunity to lay out any other procedures that will help secure public trust in such systems. If appropriate, the public authority might want to identify how individuals can appeal decisions involving algorithmic systems, to make clear what appeals processes might cover a given system's decision, or to share its mitigation strategy should the system behave in an unexpected and harmful way [445, 460]. If a harm, an undesirable outcome, or an error is identified, the public authority should explain how it intends to correct or remedy the issue, and a proposed plan for when such mitigating or corrective measures will be implemented.

Trade Secrecy

Public authorities will need to commit to accountability in both their internal technology development plans and their vendor and procurement relationships. For example, the disclosure of algorithmic systems and meaningful information about those systems will not be feasible if essential information is shielded from review by blanket claims of trade secrecy [461]. While there are certainly some core aspects of systems that have competitive commercial value, it is unlikely that these extend to information such as the existence of the system, the purpose for which it was acquired, or the results of the public authority's internal impact assessment.

Nor should trade secret claims stand as an obstacle to ensuring meaningful external research on such systems. AIAs provide an opportunity for authorities to raise any questions or concerns about trade secret claims in the pre-acquisition period, before entering into any contractual obligations. If a vendor objects to meaningful external review, this would signal a conflict between that vendor's system and public accountability. Such scenarios may require that authorities ask potential vendors to waive restrictions on information necessary for external research and review [462]. At minimum, vendors should be contractually required by authorities to waive any proprietary or trade secrecy interest in information related to accountability, such as those surrounding testing, validation, and/or verification of system performance and disparate impact [463]. This also encourages a competitive landscape among government technology vendors to meet the accountability requirements of AIAs if they want to do business with public authorities.

Meaningful access to outside researchers

AIAs should provide a comprehensive plan for giving external researchers and auditors meaningful, ongoing access to examine specific systems, to gain a fuller account of their workings, and to engage the public and affected communities in the process. This plan should give experts rapid access to a system once it is deployed (e.g. within six months). However, in situations where internal public authority assessments are insufficient or where particular risks or harms have gone unaddressed, external researchers and auditors could raise the need for pre-deployment review in the comment period. While certain individuals and communities may wish to examine the systems themselves, this cannot be relied upon: it would be unreasonable to assume that everyone has the time, knowledge, and resources for such testing and auditing [464]. Algorithmic systems can be incredibly complex, and issues like bias and systematic errors may not be easily determined through the review of systems on an individual, case-by-case basis [204]. A plan to grant meaningful access to qualified researchers would allow individuals and communities to call upon the trusted external experts best suited to examine and monitor a system to assess whether there are issues that might harm the public interest [465, 466].

To do this well, it is important to recognise that the appropriate type and level of access may vary from public authority to public authority, from system to system, and from community to community. The risks and harms at issue in different systems may demand different types of assessment and auditing using different methods and disciplines. While the right to an explanation concerning a specific automated decision could prove useful in some situations, many systems may require a group-level or community-wide analysis. For example, an explanation for a single racial profiling incident will not reveal the greater discriminatory pattern.

Many systems may only require analysis based on inputs, outputs, and simple information about the algorithms used without needing access to the underlying source code [467]. We expect that for many systems, authorities would have to provide training data or a record of past decisions to researchers. We believe that the best way for authorities to develop an appropriate research access process initially would be to work with community stakeholders and interdisciplinary researchers through the notice and comment process. Importantly, given changing technologies, the developing research field around accountability, and the shifting social and political contexts within which systems are deployed, access to a system will almost certainly need to be ongoing, and take the form of monitoring over time [468].

As an individual public authority works with researchers and community members to design its research access provisions, there are a number of elements that should be in place. Research and auditing performed on these systems should be accountable to the public, and should include a public log of which researchers and experts are provided access, and on what basis. Public authorities should ensure that affected communities are able to suggest researchers that they feel represent their interests, and should work with researchers to ensure that these communities have a voice in formulating the questions that are asked and addressed by research and auditing. Importantly, to ensure public accountability and a thriving research field, research findings and conclusions should be published openly (even if after an embargo period), and be held to standards of scrutiny and peer review within the appropriate research domains.

Ongoing auditing and research access would allow public authorities, researchers, and affected communities to work together to develop their approaches to testing and interrogating these systems. This is especially important given that the research about algorithmic accountability is young and technological development proceeds rapidly. We do not yet know what future tools, techniques, and perspectives might best keep systems accountable. External experts from a wide variety of disciplines will need the flexibility to adapt to new methods of accountability as new forms of automated decision-making emerge [469, 470].

Funding and resources issues

There is also a real danger that relying on external auditing will become an unfunded tax on researchers and the affected communities they engage with, who might be expected to take responsibility for testing and monitoring algorithmic systems without resources or compensation. Alternatively, if in-house auditors are relied on, they could become captured by the incentives of their clients or face conflict-of-interest issues. However, there are approaches that legislation could adopt to address this.

An AIA framework could fund an independent, government-wide oversight body, like an inspector general's office, to support the research, access, and community engagement [471, 472]. Community institutional review boards could be supported to help steer and review research proposals [473]. Funding could be set aside for the compensation of external auditors. Fortunately, there are many options that jurisdictions could consider for their own needs. A growing community of computer scientists, journalists, social scientists, and engaged community advocates have already proven there is an appetite for research into public algorithmic systems. This work should continue to be strongly supported by funding bodies and research authorities.

Enhanced due process mechanisms to challenge inadequate assessments or failure to mitigate

The AIA process provides a much-needed basis for evaluating and improving public authority systems. But without oversight, AIAs could become a checkbox that authorities mark off and forget, potentially sidelining community concerns [474]. That is why the AIA process should also provide a path for the public to challenge an public authority if it fails to comply with AIA requirements or if its self-assessment process was deficient in adequately identifying or addressing key concerns. For example, if an public authority fails to disclose a system that should have reasonably been considered an algorithmic system, or if it allows vendors to make overbroad trade secret claims blocking meaningful system access [475, 476], the public should have the chance to raise concerns with a public authority oversight body or directly in a court of law if the public authority refuses to rectify these problems after the public comment period. The AIA process should give the public the opportunity to effectively challenge the public authority's adoption of the system and prevent the system from being used when it fails to benefit affected communities. In Santa Clara, California, for instance, a law passed in 2016 requires the local Board of Supervisors to explicitly approve new surveillance technology before moving forward with its use [477].

Renewing Algorithmic Impact Assessments (AIAs)

In order to ensure assessments remain current and incorporate the latest information and research, authorities should be required to renew AIAs on a regular schedule. The renewed AIA will also have renewed comment and due process challenge periods. For example, authorities could be required to conduct a new AIA on all of their systems every two years. However, if there have not been significant changes to the system, to the context of its deployment, or to the need for external research access, the public authority should be allowed to minimally update their original AIA content as part of the renewal process.

Algorithmic Impact Assessments In Practice

The AIA framework goes beyond just the components described above. Those parts of an AIA — the definition of 'algorithmic system,' public authority self-assessment, public participation, and external meaningful researcher access — must be structured into a process that ensures pre-acquisition review of algorithmic systems and the opportunity for public input to be solicited and addressed.

In practice, the AIA process will not necessarily look identical between different national contexts because of existing government procurement and development practices and local interests or laws relevant to individual member states. The below description of a possible AIA process describes what a model AIA framework can look like.

1) Public authority publishes its definition of 'algorithmic system'

Public authorities should first define 'algorithmic system,' so that the public can understand how the authority decides which systems should be subject to the AIA process from the outset. This should happen before the rest of the AIA to give the public authority the opportunity to work out a definition before committing to a full review of a system that may not need it.

Once a definition is published, the authority may choose to solicit public participation on the definition alone outside of the scope of a single AIA. For example, even if the agency does not believe they have any 'algorithmic systems,' a separate process will be necessary so that the authority can still publish a definition that the public can then comment on and, if necessary, challenge. Subsequent AIAs could then use the new definition going forward.

2) Public authority publicly discloses purpose, scope, intended use and associated policies/practices, self-assessment timeline/process and potential implementation timeline of the system OR public authority publishes its decision not to review a potential system in their systems decision archive.

A public authority should first give the public basic notice of a potential new system before proceeding with the self-assessment. The public authority can use this opportunity to solicit early external feedback. This early feedback will help the public authority focus its self-assessment on the most pertinent public concerns before committing to a particular analysis.

If the public authority has decided that they do not need to conduct a full AIA, then the public authority should record that decision in their 'systems decision archive' in lieu of publishing a self-assessment timeline. This functions similarly to some existing environmental impact assessment frameworks that require authorities to first decide if their proposed action requires an assessment or is excluded from the assessment requirement [478]. Like the full self-assessment, this decision should be published before a system has entered use so that its exclusion can be challenged.

3) Public authority performs and publishes self-assessment of the system with focus on inaccuracies, bias, harms to affected communities, and describes mitigation plans for potential impacts.

The self-assessment will analyse the algorithmic system to study potential sources of bias, inaccuracies, or other harms that are of public concern. To learn about public concerns early and start to address the issues the public raises, the public agency conducting the assessment should proactively engage with the public to better scope the assessment.

4) Public authority publishes plan for meaningful, ongoing access to external researchers to review the system once it is deployed.

Ideally, stakeholders who may have helped with the public authority’s self-assessment will include the sorts of researchers who should be given ongoing access to the algorithmic system. With those stakeholders’ input, the self-assessment will likely reveal what concerns or potential problems will require ongoing monitoring and what challenges exist in studying a particular algorithmic system.

5) Public participation period

Once the full AIA has been conducted, a final period for public participation must be set aside to allow stakeholders to comment on the final product. This final public participation period allows the public to voice any concerns or issues that may have been missed during the AIA process, which the public authority should be required to address before finalizing the AIA. Ideally, the public authority will have already addressed public concerns through proactive engagement while defining 'algorithmic system,' scoping the self-assessment, and designing researcher access.

6) Final version released

The AIA should be finalised only once concerns and issues raised in the public participation period have been addressed. Any documents pertaining to the AIA should be made publicly available.

7) Renewal of AIAs on a regular timeline

Public authorities must be required to renew their Algorithmic Impact Assessments on a regular schedule. Algorithmic systems and how they are used may change over time, requiring new rounds of analysis to revisit if those changes significantly impacted how the algorithmic system operates. New concerns may come to light that were not addressed in the original AIA. Researchers might develop new techniques to analyse algorithmic systems the public authority could leverage in a future review.

8) Opportunity to challenge failure to mitigate issues raised in the public participation period or foreseeable outcomes.

Once the AIA has been finalised, the public should be given the opportunity to challenge the public authority's failure to implement mitigating or corrective measures that were raised in the AIA process. The public should also be able to challenge a public authority's decision to not conduct an AIA for a particular system if the public authority decides that the system does not meet the definition of 'algorithmic system.'

Accountability and transparency requirements for public procurement of algorithmic systems

The EU has promulgated three Public Procurement Directives that set out a framework and procedures for procurement by public authorities [479, 480, 481]. During the specification stage, contracting authorities define the requirements for the product or services they intend to procure. In 2014, the EU Public Contracts Directive expanded the scope of issues that can be covered during specification to include performance, equality, and social/environmental issues as well as any relevant processes and methods linked to these issues.

Many, if not most, algorithmic systems come into government via procurement procedures. These procedures can serve as powerful moments to raise and address transparency and accountability concerns with algorithmic system vendors. Given the expanded scope of issues that can be covered during the specification stage of procurement, contracting authorities should establish a variety of transparency and accountability requirements, particularly as they relate to performance, equality, and social issues.

In order for a public authority to adequately assess an algorithmic system's performance, it will require significant information about the system. Often authorities lack necessary information to assess performance because vendors use intellectual property claims to avoid providing necessary information or authorities fail to inquire about relevant information. There are a variety of requirements a public authority can explore to ensure it has enough information to assess the performance of an algorithmic system and ensure maintenance for optimal performance. The following recommendations seek to address performance and transparency issues.

- A public authority should require a vendor to produce materials that can be used to explain the systems in general to all stakeholders, and a technical manual including user, design and code documentation.

- The public authority should require a vendor to provide notice of any claims, lawsuits or actions related to an algorithmic system that may impact the public authority's use.
- The public authority should require a vendor to provide a comprehensive list of all proposed data sets and methodologies that the vendor intends to use in the design, production or configuration of the system. The public authority should also require the vendor to provide a datasheet assessing the quality of the data set, an explanation of any proposed manipulation of the data, and a plan to account for possible sources of bias in data collection and manipulation. Finally, the public authority should require the vendor to disclose any evidence, analysis, or reports of known or discovered flaws in the system or any relevant data sets.
- The public authority should require a vendor to produce a test version of the system and conduct a performance analysis. This performance analysis should study potential operator use of the system to understand the patterns of use and how the operator interprets or acts on the system's outputs.
- The vendor should agree to not assert any legal claims against the public authority or third parties for conducting research to test, audit, examine or otherwise understand the system's effect on an individual or group of individuals impacted by the system's outputs.

Algorithmic systems often use and produce sensitive government information. Sometimes, individuals or companies that wish to keep sensitive information private, will seek services from the private sector. To address these privacy-related equality concerns, the public authority should develop requirements that clarify ownership and confidentiality mechanisms. The following recommendations seek to address privacy, accountability, and equality issues.

The public authority should require that any materials, processes, products and related technical or use materials are exclusively owned by the public authority.

- If the vendor uses services of individuals or organisations as subcontractors, the public authority should require the vendor to use a subcontract agreement to ensure that all copyrightable work product remains the property of the public authority and that the vendor will take responsibility for the actions of its subcontractors.
- The public authority should require all materials supplied by the public authority be held in strict confidence and the public authority should prohibit copying, duplicating, disseminating or discussing public authority materials with anyone other than persons authorised by the public authority.

3.11. Development of Industry Standards for algorithmic decision-making systems

Industry standards, such as those developed by national (e.g. AFNOR (Fr), DIN (De), UNI (It), BSI (GB), ANSI (US), etc.) or international standards setting bodies (e.g. ISO, IEC, ITU, IEEE, etc.) play an important role in establishing common reference structures to enable successful industry self-regulation. The ISO 9000 family of 'Quality Management System' standards [482] for instance facilitates trust between procurers and suppliers in global supply chains. While there exist a number of standards that are related to AI/algorithmic decision-making, there currently are no international standards that deal with these directly.

In 2016 the Institute of Electrical and Electronics Engineers (IEEE) launched the IEEE Global Initiative on Ethics for Autonomous and Intelligent Systems [483] to address growing concerns about unintended consequences of algorithmic systems. Part of this initiative was the launch of the development of the IEEE P7000 series of ethics based standards, e.g. P7001 Transparency of

Autonomous systems, P7003 Algorithmic Bias Considerations. The first of these standards developments is expected to reach completion in late 2019.

At the start of 2018, the Joint Technical Committee for information systems (JTC 1) of the International Standards Organisation (ISO) and the International Electrotechnical Commission (IEC) set up a new subcommittee (ISO/IEC JTC 1/SC 42) to identify necessary standards projects for Artificial Intelligence [484]. The average development time for these international standards is three to five years. The second Study Group (i.e. SG2) that was set up in ISO/IEC JTC 1/SC 42 was tasked with investigating approaches to establish trust in AI systems through transparency, verifiability, explainability, controllability, etc. (SG1 is focusing on terminology). In September 2018 a proposal was submitted to the Artificial Intelligence subcommittee (SC42) of JTC 1, with a request to the IT Service Management and IT Governance subcommittee (SC40) to establish a joint working group to pursue work on 'Governance implication of the use of AI by organisations'.

While the current lack of established standards for algorithmic decision-making systems poses a challenge for regulatory authorities seeking references for identifying best-practices, the current stage in the standards development process provides opportunities for signalling priority areas to the Standards Setting Organisations.

3.12. Human Rights as foundation for algorithm governance

Many of the concerns that are driving the demands for algorithmic accountability are directly related to (inadvertent) violations of fundamental human rights [36, 48-54, 57, 62-64, 86, 91, 92, 93, 95]. The algorithmic systems which are automating (complex) procedural tasks, lack the capacity of human decision-makers to understand the 'human condition' and notice when a decision would infringe upon human rights, unless such a test is deliberately coded by developers into an outcomes evaluation routine. The fact that these fears are not wholly unjustified is evident from the large number of prominent examples of infringements of human rights that have been reported on, ranging from racism [e.g. 485], invasions of privacy [e.g. 486] to interference with freedom of expression [e.g. 487], restrictions of due process in criminal justice proceedings [488] and more.

While many in the technical and academic community discussing these issues have framed these problems in the language of concerns around ethical behaviour (while also referring to legal concepts of discrimination such as 'differential impact'), human-rights NGOs and researchers have started to pick up these issues as a matter of human rights. In May 2018, the leading conference on human rights in the digital age, RightsCon [489], featured a dedicated conference track on 'Artificial Intelligence, Automation and Algorithmic Accountability' and saw the launch of the 'Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems' [490]. In relation to the establishing of governance frameworks for algorithmic transparency and accountability, the key message from the Toronto Declaration is:

'As discourse around ethics and artificial intelligence continues, this Declaration aims to draw attention to the relevant and well-established framework of international human rights law and standards. These universal, binding and actionable laws and standards provide tangible means to protect individuals from discrimination, to promote inclusion, diversity and equity, and to safeguard equality. Human rights are "universal, indivisible and interdependent and interrelated". [490]

Assessment of both theoretical and practical human rights law based approaches for assessing and regulating algorithmic systems were summarised in a submissions by the Human Rights, Big Data and Technology project [491] in response to the call for evidence by the UK House of Lords Select Committee on Artificial Intelligence [492]. The starting point for this assessment was the observation that the international human rights framework is much broader than the right to privacy, freedom of expression and association, and equality and non-discrimination, and that it places a legally binding obligation on nation states to respect, protect and fulfil human rights [493]. Additionally, under the UN Guiding Principles on Business and Human Rights, businesses have a responsibility to respect human rights [493]. 'A human rights-based approach provides a system that can be applied to plans, policies and processes in order to ensure that those most centrally affected are considered and centrally involved [494, 495]'. Flowing from the obligations and responsibilities imposed by human rights law, '[a] human rights-based approach offers increased transparency within policy formulations, and 'empowers people and communities to hold those who have a duty to act accountable' [496]. At a practical level, McGregor [491] presents the following as 'necessary, specific and unique' to a human rights-based approach:

1. 'Assessment and analysis in order to identify the human rights claims of rights-holders and the corresponding human rights obligations of duty-bearers, as well as the immediate, underlying, and structural causes of the non-realisation of rights.
2. [Assessment of the] capacity of rights-holders to claim their rights, and of duty-bearers to fulfil their obligations, [followed by the development of] strategies to build these capacities.
3. [Monitoring and evaluating] both outcomes and processes guided by human rights standards and principles.
4. [Programming, processes, policies and planning are] informed by the recommendations of international human rights bodies and mechanisms'. [497]

To illustrate how a human rights-based approach to accountability would apply in the realm of algorithmic decision-making, McGregor [491] described how: '[i]n this context, a human rights-based approach requires using international human rights standards and norms as a means for identifying and defining elements within the algorithm life-cycle that give rise to human rights concerns, establishing which entity/entities impact(s) upon which rights, addressing questions of responsibility, and identifying how human rights concerns can be addressed'.

A human rights-based approach to the accountability of algorithmic decision-making, such as proposed by McGregor [491] would include tools such as initial and ongoing human rights impact assessments to test and review the impact of algorithmic decision-making on human rights.

'Human rights impact assessments [498] are 'instruments for examining policies, legislation, programmes and projects prior to their adoption to identify and measure their impact on human rights…They are designed to identify the intended and unintended impact on the enjoyment of human rights, and the State's ability to protect and fulfil them. As such, they are a planning tool to prevent human rights violations by assessing the formal or apparent compatibility of laws, policies, budgets and other measures with human rights obligations, as well as the likely impact in practice, thus creating the opportunity for reconsideration, revision or adjustment prior to adoption'. [491]

3.13. Global dimension to algorithm governance

One of the defining characteristics of the digital economy and the technologies it is based on, including internet platforms, cloud computing and algorithmic systems is the high degree of cross-boarded and global reach of the services that are built on these technologies. To successfully govern these technologies therefore requires a global dialogue and collaboration across borders- among both rich and poor countries, to avoid a patchwork of country-specific or regional approaches. In this section we review some of the international dimensions of governing algorithmic transparency and accountability.

Geopolitical competition

The introduction of algorithmic processes to increase machine autonomy and automate much of the services sector is now globally accepted to represent a significant shift in society akin to a '4th

Industrial Revolution' [499]. This perception has raised the spectre of the sharp rise in economic, military and political power of those countries that managed to be the first to industrialise in the 18th century. Many nations, including the US, China and various EU states have responded to this view by publishing ambitious 'National Artificial Intelligence Strategies' [500 intended to ensure that they will be among the leaders and winners of this new industrial revolution. This has led some commentators to raise alarm over a global 'AI arms race' [e.g. 501, 502, 503] with special concerns regarding the dual-use (civilian and military) nature of most algorithmic methods [e.g. 504, 505] and their potential application in cyber-operations [506] and Autonomous Weapons Systems [507]. In such a competitive environment with a strong 'winner-takes-all' narrative [508] there are strong pressures to push for computational efficiency and functional performance of algorithmic systems at the cost of non-functional considerations (i.e. considerations that do not directly contribute to the ability of the system to perform its task) such as transparency [509]. Within this hyper-competitive environment, any regulatory intervention to mandate algorithmic transparency is likely to be met with similar suspicions of protectionist interventionism as has been the case with the GDPR [510, 511]. As a counter point however, the GDPR has also inspired various countries to either enact similar legislation, such as Brazil [512], China [513], India [514], South Africa [515], California [516], etc. This has further increased their motivation to join the Council of Europe's Convention 108 (e.g. Senegal, Mauritius, Tunisia, Cabo Verde, Mexico [517]), which in its modernised form [518] is closely aligned with the GDPR and represents the most viable basis for a truly global data privacy framework [519]. The example presented by the GDPR suggests that regulatory interventions geared towards strong protections of citizen rights can position the EU as a viable leader that many states are willing to engage with.

Racial and cultural bias in algorithms

Most algorithmic systems are created by relatively homogenous groups of developers [520] using data sets that frequently over-represent some groups while under-representing others. The ImageNet data set for instance, which is a core data set in the creation of computer vision applications, is populated for more than 45% with images from the United States, whereas images from China and India together contribute just 3% of the date. This lack of diversity is partially responsible for failures of image recognition algorithms that interpret Asians eyes as always blinking; are capable of labelling a photograph of a traditional US bride dressed in white as 'bride', 'dress', 'woman', 'wedding', but a photograph of a North Indian bride as 'performance art' and 'costume'; and misclassify darker-skinned women's gender with an error rate of 35% while lighter-skinned men are misclassified at a rate of only 0.8% [521]. While such biases can be relatively easy to notice in the case of image recognition systems, other algorithms that are more deeply embedded within services may contain severe cultural biases that are difficult to detect without access to knowledge about how the system works. An illustrative example of such a bias was brought to light in 2016 when it was revealed that the database used by the Facebook Trending Topics app (since discontinued), which selected news items to recommend to readers, consisted of 1000 trusted news sources with many of the world's major news outlets missing from the list and many countries, especially in Africa and eastern Europe, not having even a single outlet listed. As a result, despite being active in countries around the globe, the recommendations by the Facebook Trending Topics app were heavily skewed towards news items reported in Anglophone media [522]. In order to limit unintended cultural and/or racial discrimination by algorithmic systems when they are deployed in a different societal/cultural context than where they were developed, it may be necessary to obtain transparency reports that document how the development team has addressed the societal/cultural localisation challenges.

Foreign interference

While the usually unintentional types of algorithmic bias discussed in the previous paragraph can introduce undesirable cultural interference, a more sinister and deliberate form of foreign interference is represented by:

1. The targeted use of algorithmic systems, such as social media bots [e.g. 523, 524] and big data analytics for psychological micro-targeting of political ads [e.g. 525, 526], to intervene in the informational integrity of national electoral processes [527].
2. Offensive cyber operations [528] for (corporate) espionage [e.g. 529], disruption of national services [e.g. 530] and/or interference in elections [e.g. 531, 532].

To counter each of these forms of foreign interference it is vital to establish highly confident attribution mechanisms in order to hold the true offending party to account. While national counter-intelligence work is beyond the scope of this document, it is worth noting the important role that transnational bodies such as Europol's Cyber Crime units [533] and NATO's Cooperative Cyber Defence Centre of Excellence [534] play in vital information coordination in response to cyber incidents,

As was shown by the Cambridge Analytica case [360] (see also section 3.4.1), a further important element for combating foreign interference is the ability to obtain cooperation, or when necessary legally compel, transparency and accountability from corporate actors (e.g. Facebook in the case of Cambridge Analytica incidents) whose algorithmic systems were involved in information/cyber interference operations [535].

Trade Negotiations

In the context of the global dynamics of geopolitical interference, algorithmic cultural bias and the use of algorithmic systems for foreign interference, it is important to note that current e-commerce proposals being discussed at the WTO (as well as regional trade negotiations such as the Regional Comprehensive Economic Partnership (RCEP) and others) include proposals to protect Intellectual Property by restricting access to information regarding proprietary algorithms. The Japanese proposal at the WTO for exploratory work on an electronic commerce initiative [536] for instance includes a clause on 'Prohibition of Disclosure of Important Information such as Trade Secrets Including Source Codes and Proprietary Algorithms.' A similar statement on an electronic commerce initiative that was filed on the same day by the United States [537] also includes a clause on protection of proprietary information, with sections on Protecting Source Code, Barring Forced Technology Transfer and Barring Discriminatory Technology Requirements. The degree to which similar clauses in free trade agreements might cause problems for accountability and regulatory oversight of algorithmic systems will depend on the details of the agreements that are finally produced. It will be necessary to find a workable balancing point between trade secrets and transparency, similar to policies in other domains [538]. Agreements may need to indicate what factors or metrics of the algorithm would be disclosed, the frequency of their disclosure (e.g., daily, monthly, or real-time), and the vehicle for communicating such information (e.g., a separate document, or integrated into the algorithmic output in some way).[204]

International coordination in algorithm governance

As discussed earlier in this section, many aspects of algorithmic accountability involve global transnational interactions between states and globally operating corporate actors. In order to structurally address issues of transparency and accountability of algorithmic systems will therefore require ongoing coordination at an international level. Current international efforts to deal with issues arising from the use of algorithmic systems are highly dispersed across multiple sector-specific initiatives such as UN led efforts to:

- Encourage the development of 'AI for Good' to address the UN Sustainable Development Goals (SDGs), managed by ITU. [539]
- Ban Lethal Autonomous Weapon Systems, primarily discussed under the banner of the UN Convention on certain Conventional Weapons (UNCCW). [540]
- Multi-stakeholder dialog on ethics and governance of AI/algorithmic systems, in the frame of the Internet Governance Forum (IGF) under the umbrella of UNESCO. [541]

But also efforts by the Council of Europe (e.g. Committee of experts on Internet Intermediaries (MSI-NET) [542]), the Innovation Ministers of the G7 [543] and G20 [544] and the OECD [545]. At the same time, bilateral and multilateral trade negotiations are asserting the importance of intellectual property rights and trade secrets relative to transparency and accountability.

In order to establish consistent international governance of algorithmic systems, and establish a cooperative alternative to the winner-takes-all arms race narrative it may be necessary to establish a new international body, possibly within a UN agency. Such a body could help to coordinate national regulations by establishing common interests and values for accountable use of algorithmic systems, drawing on existing international human rights standards and norms to provide enhanced certainty and ensure international perspectives that are based on universal values [e.g. 546, 547].

Though still at an early stage of development, some efforts towards the establishment of trans-national coordination on governance of algorithmic systems/AI is starting to emerge at fora such as the OECD's Artificial Intelligence Expert Group (AIGO) [548], the World Government Summit's Global Governance of AI Roundtable [549], and activities on AI at the Council of Europe [550].

Some preliminary analyses on the requirements and potential frameworks for a global coordination forum for AI governance have been explored in recent publications by Erdelyi and Goldsmith [551], and Wendell and Marchant [552]. Erdelyi and Goldsmith propose an International Artificial Intelligence Organization (IAIO), which would serve as international forum for discussion as well as international standards setting, similar to the role of the ITU for telecommunications [551]. The paper by Wendell and Marchant proposes an approach framed around Global Coordinating committees (GCCs) in which an international GCC would work with complementary regional bodies to reinforce the governance initiatives of organizations such as the IEEE, WEF, Partnership on AI and various research centres [552].

4. Policy options

Based on our review and analysis of the current literature regarding algorithmic transparency and accountability, and the successes, failures and challenges of different governance frameworks that have been applied to technological developments (especially in ICT), we propose a set of four policy options each of which addresses a different aspect of algorithmic transparency and accountability:

1. Awareness raising: education, watchdogs and whistleblowers.
2. Accountability in public sector use of algorithmic decision-making.
3. Regulatory oversight and Legal liability on private sector.
4. Global dimension of algorithmic governance.

4.1. Awareness raising: education, watchdogs and whistleblowers

Over a decade of struggle with consent based approaches to data privacy have shown how information asymmetries between service providers and consumers have limited the ability of citizens to successfully exercise their rights when interacting with digital services [553]. When it comes to algorithmic decision-making, the prevailing consensus in the literature suggests that consumer/citizens are struggling to understand how these systems work, the impact they are having, and how to critically evaluate their decisions [554, 555, 556, 557, 558, 559, 560, 561]. The same is true for many highly skilled non-technical professionals, e.g. judges and lawyers [562, 563, 564]. This lack of 'algorithmic literacy' is limiting the ability of people to express agency in their interaction with these systems, and thereby undermining the functioning of demand side market pressure to self-regulate the sector. In order for algorithmic transparency to enable accountability [340, 341, 342]. Any general understanding of algorithmic functioning, however will do little to provide accountability unless it is combined with some form of public disclosure about the types and properties of the algorithms (and data) associated with a specific decision. In order to be useful such notifications should be standardised and short, akin to nutrition labels [20, 346, 347] or restaurant inspection scores [204]. Information included in a 'disclosure label' should be limited to that which has the potential to either impact an individual user's decision processes, or wider public understanding of aggregate system behaviour [340]. Beyond helping citizens to navigate their personal interactions with algorithmic decision systems, 'algorithmic literacy' is also important to help the understand media reports related to algorithmic decisions, and participate in the public dialog about the use of these system.

Investigative journalism and whistleblowers play an important role in uncovering questionable uses and outcomes of algorithmic decision-making, and challenging the lack of accountability. Clear examples being the Cambridge Analytica case [360], revelations by Edward Snowden regarding questionable reliability of algorithmic targeting of drone strikes [565], the controversy around the COMPAS 'recidivism algorithm' used in various US courts [94] and many more [e.g. 361, 362, 363, 566], including the importance of whistleblowing (through the media) as part of (ex-)employee led activism aimed at changing company projects that are perceived to be unethical [245, 246, 249, 250, 253] (see also appendix 1 on examples of public scrutiny much of which was journalist led). A New York Times investigation revealed that ride sharing company Uber used an algorithm to flag and evade regulators in cities all over the world. Journalists learned about the algorithm's existence and purpose by speaking with current and former Uber employees and reviewing documents these sources provided [567]. The Times' investigation led to broad media coverage and a Department of Justice inquiry into potential criminal behaviour by the company [568].

Beyond their role as independent watchdogs, journalists help to present relevant aspects of algorithms to the wider audience in plain language with understandable narratives. Several of the journalistic investigations listed above have sparked broad public conversations and important normative debates, including triggering series of academic studies [205]. The Propublica report on 'Machine Bias' in the COMPAS algorithm [93] for instance triggered a series of studies into the meaning of 'fair' and 'unbiased' algorithmic systems [e.g. 569, 570, 571] and the impossibility of producing a system that would simultaneously be unbiased on measures of 'overall misclassification', 'False Positive rates' and 'False Negative rates' [572].

In order to uncover cases of algorithmic 'malpractice' journalists are combining interviews, right to information requests, and investigative reporting, with computationally intensive methods for reverse engineering algorithms (e.g. 'black box testing') [94, 573], which has developed into a small but active field of 'algorithmic accountability journalism' that grew out of the more established field of 'data journalism' [204]. The reverse engineering process focuses on the system's performance in-use and can therefore tease out consequences that might not be apparent even when the journalist speak directly to the designers of the algorithm. Legally however, this reverse engineering of commercial software is prohibited by Trade Secrets and the Copyright (e.g. the DMCA in the US). Software vendors also typically add anti-reverse engineering clauses to End User License Agreements (EULAs),[574] forcing the decision: Is it okay to breach such a contract if it gets you closer to the truth about the algorithm? This raises the need for establishing exemption clauses for public interest reverse engineering of software by 'algorithmic accountability journalists. Something similar to the EU Whistleblower Protection directive that was proposed earlier this year [575].

Developing the skills to do algorithmic-accountability reporting takes dedicated efforts to learn the computational thinking, programming, and technical skills needed to make sense of algorithmic decisions. While there is growing awareness of more complex algorithms among data journalists, the number of computational journalists with the technical skills to do a deep investigation of algorithms is still rather limited [204]. Supporting computationally literate reporters by providing computational infrastructure and tech-savvy computer scientists for then to team up with would help to facilitate quality algorithmic accountability reporting. Another way would be to provide support (e.g. scholarships and dedicated courses) to train journalists themselves in more computational techniques.

Besides technology skills and the legality of reverse engineering, investigative reporting on algorithms also requires an understanding of the ethical questions that arise from the possible ramifications of publishing details of how certain algorithms work. Would publishing such information negatively affect any individuals? By publishing details of how an algorithm functions, specifically information about what inputs it pays attention to, how it uses various criteria in a ranking, or what criteria it uses to censor, how might that allow the algorithm to be manipulated or circumvented? Who would benefit from that manipulation? [204] In order to help investigative journalism of algorithms perform their watchdog function, while minimising negative side-effects there should be financial and logistical support for coordinated, fact-checked and vetted algorithms journalism, similar to the collaborative journalism efforts behind the publications of the Snowden files, the Panama Papers and the Paradise Papers.

Recommendations:

In order for people to have agency and be able to critically evaluate the results they are given by algorithmic systems, they must have a basic understanding of how algorithmic decision-making works. We therefore recommend:

- The provision of 'algorithmic literacy' that teaches core concepts such as: computational thinking, the role of data and the importance of optimisation criteria.
- The introduction of standardised notification practices to communicate the type and degree of algorithmic processing involved in decisions.

In order for democratic society to function, those in power (political or otherwise) must be held accountable. Much of the critical discourse on the use and abuse of algorithmic decision-making relies on investigative reporting and whistleblowers to identify the existence of issues such as algorithmic bias, manipulation and surveillance etc. We therefore recommend:

- The provision of computational infrastructure and access to technical experts to support the data analysis and algorithm reverse engineering efforts of 'algorithmic accountability journalists'.
- Whistleblower protection (expanding the current EC proposal to include any violation of human rights) and protection against prosecution on grounds of breaching copyright or Terms of Service when doing so served the public interest.

4.2. Accountability in public sector use of algorithmic decision-making

Algorithmic systems are increasing being used by public authorities to improve efficiencies, implement complex processes and support evidence-based policy making. Due to the nature of public sector responsibilities, these uses of algorithmic systems have potentially far reaching impacts sometimes involving the weakest members of society. The use of algorithmic systems in public services therefore requires extra levels of transparency and accountability. Public sector procurement is also a major source of business for many companies and as such provides a route for incentivising commercial development of transparent and accountably systems. We therefore recommend Algorithmic Impact Assessments as part of public sector use and procurement of algorithmic systems.

Algorithmic Impact Assessments is a framework designed to help policymakers and their constituents understand where algorithmic systems are used within government, assess the intended use and proposed implementation, and allow community members and researchers to raise concerns that require mitigation. This framework draws on the history and development of assessments in other areas such as environmental policy, privacy law, and data protection. It also builds on growing and important research on algorithmic accountability. The framework requires public authorities to perform a self-assessment of the algorithmic systems it intends to use and they will likely require additional information from vendors in order to perform this assessment adequately. In practice the exact steps of an Algorithmic Impact Assessment (AIA) is likely to depend on the national context and the sensitivities of the specific public sector branch, however the general shape of the process is likely to include the following:

- *Publication of public authority's definition of 'algorithmic system'.* This allows the public to understand how the authority decides which systems will be subjected to AIAs. This definition must be periodically reviewed and should involve public participation.

Once the definition has been published and gone through public review, it is used to asses all currently used systems, and any bid in response to a tender for procurement of an 'algorithmic system' by the public authority.

1. *Public disclosure of purpose, scope, intended use and associated policies/practices, self-assessment timeline/process and potential implementation timeline of the algorithmic system OR publication (and archiving) of the decision not to review a potential system.* Publication at the start of the assessment process provides opportunity for early external feedback which can help to focus the assessment on the most pertinent public concerns.
2. *Performing and publishing of self-assessment of the system with focus on inaccuracies, bias, harms to affected communities, and describes mitigation plans for potential impacts.* This

should include proactive engagement with the public who will be most affected by the intended use of the system, in order to better scope the assessment.

3. *Publication of plan for meaningful, ongoing access to external researchers to review the system once it is deployed.* Even though the AIA attempts to anticipate and mitigate potential negative impacts of introducing a algorithmic decision system, it is likely that not all effects will be (correctly) anticipated. It is therefore important to make sure that the system is set up in a way that facilitates external monitoring at an ongoing basis.

4. *Public participation period.* Once the evidence from the self-assessment of the system has been collected it needs to be communicated to the public in an understandable way and the public has to be given a change to voice their concerns.

5. *Publication of final Algorithmic Impact Assessment, once issues raise in public participation have been addressed.* Any documents pertaining to the AIA must be made publicly available.

6. *Renewal of AIAs on a regular timeline.* Algorithmic systems and how they are used may change over time, requiring new rounds of analysis to revisit if those changes significantly impacted how the algorithmic system operates. New concerns may come to light that were not addressed in the original AIA. Researchers might develop new techniques to analyse algorithmic systems the public authority could leverage in a future review.

7. *Opportunity for public to challenge failure to mitigate issues raised in the public participation period or foreseeable outcomes.* Once the AIA has been finalised, the public should be given the opportunity to challenge the public authority's failure to implement mitigating or corrective measures that were raised in the AIA process. The public should also be able to challenge the decision when the public authority decides that a system does not meet the 'algorithmic system' criteria it has set itself for triggering AIAs.

We therefore recommend:

- Member states adopt Algorithmic Impact Assessments following the process outlined in this report.
- Member states should work with public authorities to ensure that each agency develops a meaningful public education and engagement process to ensure all stakeholders can participate in the Algorithmic Impact Assessment. We recommend member states review existing recommendations on public engagement including the recommendations that were recently submitted to New York City's Automated Decision Systems Task Force, include in Appendix II.
- Member states must develop and implement accountability and transparency procurement requirements for the acquisition of algorithmic systems. Such requirements will allow public authorities to have access to necessary technical and other information that will allow the authorities to perform a robust Algorithmic Impact Assessment.

4.3. Regulatory oversight and Legal liability

Commercial development and application of algorithmic decision-making systems is undergoing rapid growth, with at times uncertain implications for citizens and society. Industry standards for best practice largely do not yet exist in this space. The interpretation of existing laws is sometimes uncertain when applied to algorithmic decision-making, and judicial experience in this context is in short supply. While much of this is simply the result of rapid dynamic growth, this must not be allowed to limit the rights and legal protections that citizens (and business customers) are entitled to.

One approach to protecting citizens from negative impacts arising from algorithmic decision-making might be to impose on all private sector uses of such systems a similar Algorithmic Impacts Assessment regime as we are recommending for public sector authorities. While doing so might make sense for high-impact applications, such as autonomous passenger vehicles or political elections related services etc., for most private sector applications the financial and administrative burden of such a requirement would not be proportionate to the risks. For low-risk uses of algorithmic decision-making, defined to a large extent by the reverse-ability of the algorithmic decision and the non-permanence of its impacts, it would be preferable to establish a legal liability framework that allows service providers to accept greater tort liability in exchange for reduced transparency and Algorithmic Impact Assessment requirements.

In order to facilitate such a tiered regulatory regime, it would be necessary to establish a specialised regulatory body with expertise in analysing algorithmic decision-making systems and a network of external expert advisors. The primary tasks of the regulatory body would be:

1. Establishing (and keep updated) a 'threat matrix' [e.g. 576] for assessing the level of regulatory oversight that is necessary for an algorithmic decision system. This should be based on factors such as: impact of its outputs (human rights implications; scale of use; (ir)reversibility of the consequences; etc.); application domain; verify-ability of its behaviour (including failure modes); explainability of decision outcomes; transparency of processing; etc.

2. Coordinating with existing domain regulators, e.g. Data Protection Authorities, Consumer Protection authorities etc., regarding application of existing laws when products/services involve the use of algorithmic decision processes.

3. Coordinating with Standards Setting Organisations (e.g. ISO/IEC, IEEE), industry and civil-society organisations to identify relevant standards and best-practices procedures that could be used for third-party certification. For some algorithmic systems that have been assessed as requiring higher levels of regulatory oversight (but not quite requiring Algorithmic Impact Assessments), such certification could become a mandatory requirement, similar to CE certification. For less critical systems the certification could serve to communicate system trustworthiness to end users and reduce the tort liability of the service/product provider.

4. Facilitating the effectiveness of the tort liability mechanism as means for regulating accountability of algorithmic systems by providing a contact point for citizens who are not familiar with legal procedures.

5. Auditing the Algorithmic Impact Assessments of high-level impact systems to approve or reject the proposed uses of algorithmic decision-making in highly sensitive and/or safety-critical application domains (e.g. private health-care). The Algorithmic Impact Assessment for private sector applications could follow a very similar process as the one we proposed for the public sector, with the possible difference that the various stages of public disclosure could be handled as confidential communication to the regulatory body (under non-disclosure agreement) in order to safeguard vital trade secrets.

6. Investigating suspected cases of rights violations by algorithmic decision-making systems, for both individual decision instances (e.g. singular aberrant outcomes) and statistical decision patterns (e.g. discriminatory bias). Investigations could be triggered following the lodging of complaints, or on the basis of evidence provided by whistleblowers, investigative journalists or independent researchers (including NGOs and academics).

We therefore recommend:

- The creation of a regulatory body for algorithmic decision-making tasked with:
    - Establishing a risk assessment matrix for classifying algorithm types and application domains according to potential for significant negative impact on citizens.
    - Investigating the use of algorithmic systems where there is a suspicion (e.g. evidence provided by a whistleblower) of infringement of human rights.
    - Advising other regulatory bodies regarding algorithmic systems as they apply to the remit of those agencies.
- That systems classified as causing potentially severe non-reversible impact be required to produce an Algorithmic Impact Assessment, similar to public sector applications.
- That systems with medium severity non-reversible impacts require the service provider to accept strict tort liability, with a possibility of reducing the liability by having the system certified as compliant with (as yet to be determined) best-practice standards.

4.4. Global coordination for algorithmic governance

As with much of the digital economy, the use of algorithmic systems is characterised by a high degree of cross-border and global reach of the services that are built on these technologies. To successfully govern algorithmic systems therefore requires global dialogue and collaboration across borders and among rich and poor countries to avoid a patchwork of country-specific or regional approaches. The narrative of a '4th Industrial Revolution' with winner-takes-all dynamics however has triggered what some have refer to as an 'AI arms race' [e.g. 501, 502, 503]. In such a hyper-competitive environment there are strong pressures to push for computational efficiency and functional performance of algorithmic systems at the cost of non-functional considerations (i.e. considerations that do not directly contribute to the ability of the system to perform its task) such as transparency [509]. Without multilateral negotiation there is a risk that under these competitive conditions any regulatory intervention to mandate algorithmic transparency may be interpreted as protectionist interventionism intended to block market access by foreign companies. In the context of these global dynamics, proposals for new e-commerce trade agreement are being discussed at the WTO (as well as regional trade negotiations such as the Regional Comprehensive Economic Partnership (RCEP) and others), which include clauses for Intellectual Property protection that would restrict access to information regarding proprietary algorithms. While the details of such restrictions remain to be determined, due care will be required to ensure that such clauses in free trade agreements do not cause problems for accountability and regulatory oversight of algorithmic systems.

International tensions also arise from the use of algorithmic systems for social media bots [e.g. 523, 524], micro-targeting of political ads [e.g. 525, 526] and other interventions in the informational integrity of national electoral processes [527], as well as offensive cyber operations [e.g. 528, 529, 530, 531, 532]. In order to effectively respond to such interference without resorting to bilateral escalation of cyber operations, it is important to have a broad international community involved in publicly establishing methods and guidelines around attribution of such attacks and defining of proportionate responses.

Though still at an early stage of development, some efforts towards the establishment of trans-national coordination on governance of algorithmic systems/AI is starting to emerge at fora such as the OECD's Artificial Intelligence Expert Group (AIGO) [548], the World Government Summit's Global Governance of AI Roundtable [549], and activities on AI at the Council of Europe [550].

Building on the international recognition as leader in data privacy legislation that the EU has established through the introduction of the GDPR, which is exemplified by the various GDPR inspired national privacy laws that are being implemented globally, the EU is currently uniquely positioned to take the lead in establishing a new international body to help coordinate approaches to algorithmic transparency and accountability at a global level.

Some preliminary analyses on the requirements and potential frameworks for a global coordination forum for AI governance have been explored in recent publications by Erdelyi and Goldsmith [551], and Wendell and Marchant [552]. Erdelyi and Goldsmith propose an International Artificial Intelligence Organization (IAIO), which would serve as international forum for discussion as well as international standards setting, similar to the role of the ITU for telecommunications [553]. The paper by Wendell and Marchant proposes an approach framed around Global Coordinating committees (GCCs) in which an international GCC would work with complementary regional bodies to reinforce the governance initiatives of organizations such as the IEEE, WEF, Partnership on AI and various research centres [552].

We therefore recommend:

- The establishment of a permanent global Algorithm Governance Forum (AGF) for multi-stakeholder dialog and policy expertise related to algorithmic systems, and associated technologies. Based on the principles of Responsible Research and Innovation, the AGF would provide a forum for coordination and exchanging of governance best-practices related to algorithmic decision-making.
- The adoption of a strong position in trade negotiations to protect regulatory ability to investigate algorithmic systems and hold parties accountable for violations of European laws and human rights.

5. Conclusions

Algorithmic decision-making systems play an increasingly important part in public and private sector decision-making processes with potentially significant consequences for individuals, organisations and societies as a whole. When used appropriately, with due care and analysis of their impacts on people's lives, algorithmic systems, including AI and machine learning, have great potential to improve the quality and efficiency of products and services. In order to achieve this, however, it is vitally necessary to establish clear governance frameworks for transparency and accountability to make sure that the risk and benefits are equitably distributed in a way that does not unduly burden or benefit particular sectors of society. Within this context transparency and accountability are both tools to promote fair algorithmic decisions by providing the foundations for obtaining recourse to meaningful explanation, correction, or ways to ascertain faults that could bring about compensatory processes.

At the level of technical properties of algorithmic systems, there has been a great deal of discourse regarding the inscrutability of 'black box' algorithms, especially in relation to machine learning systems. It is true that the complexity of the algorithmic processing, combined with the scale and variety of data involved in the computations, makes transparency in the sense of 'explaining the steps of the algorithm' unlikely to lead directly to an informative outcome. This is especially true when the system also involved the use of machine learning methods for inferring statistical models directly from the data. There are, however, technical methods for reducing algorithmic opacity, or extracting explanations for the system's behaviour despite a lack of transparency. Understanding the overall system, and understanding a particular outcome may however require quite different approaches. A key idea to keep in mind is the goal of transparency. Is the aim to understand how the system works or how it behaves? For an understanding of the overall system, i.e. to obtain a general understanding of the algorithmic decision-making process, approaches include: design/code review; input data analysis; statistical analysis of outcomes; analysis of sensitivity to inputs. One challenge with these approaches is that they are likely to be difficult or impossible without direct involvement of system developers. Understanding how a system works is likely of little value for the transparency of individual outcomes. In that case approaches providing explanation become more important. Meaningful transparency into *how* outcomes are reached is technically challenging given modern computing systems; regulatory requirements for such transparency may significantly limit the ability to use advanced computing techniques for regulated purposes. Meaningful transparency into the *behaviour* of computing systems is feasible, and can provide important benefits. Mechanisms for behavioural transparency may need to be designed into systems, and typically require the participation of system developers or operators. Algorithmic accountability, such as redress in cases of unfair treatment, are beyond the technical challenges of the algorithm. These are more a question of the actions implied by the specifications and the organisational structure surrounding the algorithmic system. While accountability for actions taken by algorithmic systems may need to be different than for human actions, those differences are largely governed by the particular application.

At a high level, the fast pace of developments and the absence of clearly established best practice technology standards suggests that a governance framework following a flexible principles-based approach is likely to provide a better balance than a rules-based approach for protecting the fundamental rights of citizens while retaining freedom to innovate new algorithmic methods.

Based on a systematic consideration of each of five governance categories (demand-side market solutions; supply-side market solutions; companies' self organisation; branches' self-regulation; co-regulation; and state intervention) it is concluded that the dynamics of the digital economy do not currently lend themselves to demand-side or supply-side market solutions for improving algorithmic transparency or accountability. While there are efforts at self organisation relating to the ethics of algorithmic decision making, there is currently no convincing evidence that these will be sufficient to provide the necessary safeguards for citizens. Branch self-regulation, or possibly co-regulation through supervisory involvement from regulatory agencies, is starting to take shape in this sector but will require the establishment of industry standards that are currently still in development. Schemes for certifying that algorithmic decision systems do not exhibit unjustified bias are also being developed. Overall, however, a role can be found for state intervention to guide innovation towards a greater focus on transparent, explainable and accountable methods. The type and level of state intervention should be carefully adjusted to match the algorithmic accountability required within the context of the good, harm and risk these systems bring. Despite problems with consent mechanisms for data privacy, information measures including algorithmic literacy education, 'food label' style notification of algorithmic decisions and scrutiny by investigative journalism (including whistleblowers) remain vital elements for delivering democratic agency to the citizen. Funding incentives for research into explainable decision-making algorithms, as well as investigations into the impact of algorithmic systems on society would also provide important stimuli for greater algorithmic accountability. Probably one of the most important direct incentives related to financial incentives is the role of public service procurement of algorithmic systems. Here there is strong potential to push for greater transparency and accountability by introducing measures such as algorithmic impact assessment requirements for systems that have a significant impact on the public. Such requirements are in keeping with the democratic responsibilities of public service provision. At the level of direct intervention through legislative means or regulatory bodies there is potential for an approach that combines risk assessment by a regulatory body with corresponding levels of tort liability.

As a final element in the assessment of governance frameworks for algorithmic systems the global nature of developments in algorithmic decision-making and the need for global coordination are highlighted.

In order to implement these assessments, four mutually reinforcing policy options are proposed, addressing:

1. awareness raising: education, watchdogs and whistleblowers;
2. accountability in public sector use of algorithmic decision-making;
3. regulatory oversight and legal liability of the private sector; and
4. the global dimension of algorithmic governance.

References

[1] Burrell, Jenna. 'How the machine 'thinks': Understanding opacity in machine learning algorithms.' Big Data & Society 3, no. 1 (2016): 2053951715622512.

[2] Dutton, David M., and Gerard V. Conroy. 'A review of machine learning.' The knowledge engineering review 12, no. 4 (1997): 341-367.

[3] Kinsella, Clare, and John McGarry. 'Computer says no: technology and accountability in policing traffic stops'. Crime, law and social change 55, no. 2-3 (2011): 167-184.

[4] Wihlborg, Elin, Hannu Larsson, and Karin Hedström. '' The Computer Says No!'--A Case Study on Automated Decision-Making in Public Authorities.' In System Sciences (HICSS), 2016 49th Hawaii International Conference on, pp. 2903-2912. IEEE, 2016.

[5] Lahlou, Saadi, Marc Langheinrich, and Carsten Röcker. 'Privacy and trust issues with invisible computers'. Communications of the ACM 48, no. 3 (2005): 59-60.

[6] Schneier, Bruce. Data and Goliath: The hidden battles to collect your data and control your world. WW Norton & Company, 2015.

[7] Medaglia, Carlo Maria, and Alexandru Serbanati. 'An overview of privacy and security issues in the internet of things'. In The Internet of Things, pp. 389-395. Springer, New York, NY, 2010.

[8] Martínez-Plumed, Fernando, Shahar Avin, Miles Brundage, Allan Dafoe, Sean Ó. hÉigeartaigh, and José Hernández-Orallo. 'Accounting for the Neglected Dimensions of AI Progress'. arXiv preprint arXiv:1806.00610 (2018).

[9] Galliers, Robert D., Sue Newell, G. Shanks, and Heikki Topi. 'Datification and its human, organisational and societal effects: The strategic opportunities and challenges of algorithmic decision-making.' (2017): 185-190.

[10] Cohen, Nicole S. 'The valorisation of surveillance: Towards a political economy of Facebook.' Democratic Communiqué 22, no. 1 (2008): 5.

[11] Lindh, Maria, and Jan Nolin. 'Information we collect: Surveillance and privacy in the implementation of Google Apps for Education'. European Educational Research Journal 15, no. 6 (2016): 644-663.

[12] Sculley, D., Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 'Winner's Curse? On Pace, Progress, and Empirical Rigor.' (2018).

[13] Baum, Seth D., Ben Goertzel, and Ted G. Goertzel. 'How long until human-level AI? Results from an expert assessment'. Technological Forecasting and Social Change 78, no. 1 (2011): 185-195.

[14] Backer, Larry Catá. 'And an Algorithm to Bind Them All? Social Credit, Data Driven Governance, and the Emergence of an Operating System for Global Normative Orders'. (2018).

[15] Zambonelli, Franco, Flora Salim, Seng W. Loke, Wolfgang De Meuter, and Salil Kanhere. 'Algorithmic Governance in Smart Cities: The Conundrum and the Potential of Pervasive Computing Solutions.' IEEE Technology and Society Magazine 37, no. 2 (2018): 80-87.

[16] Yeung, Karen. 'Algorithmic regulation: a critical interrogation.' Regulation & Governance (2017). https://doi.org/10.1111/rego.12158

[17] Velázquez, E., M. Yazdani, and Pablo Suárez-Serrato. 'Socialbots supporting human rights.' arXiv preprint arXiv:1710.11346 (2017).

[18] Colaresi, Michael, and Zuhaib Mahmood. 'Do the robot: Lessons from machine learning to improve conflict forecasting.' Journal of Peace Research 54, no. 2 (2017): 193-214.

[19] Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 'Fair, Transparent, and Accountable Algorithmic Decision-making Processes.' Philosophy & Technology (2017): 1-17.

[20] Ananny, Mike, and Kate Crawford. 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability.' New Media & Society 20, no. 3 (2018): 973-989.

[21] Felici, Massimo, Theofrastos Koulouris, and Siani Pearson. 'Accountability for data governance in cloud ecosystems.' In Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on, vol. 2, pp. 327-332. IEEE, 2013.

[22] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 'Glossary for Discussion of Ethics of Autonomous and Intelligent Systems, Version 1 [accessed on 28 September 2018] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/eadv2_glossary.pdf

[23] Pasquale, Frank. The black box society: The secret algorithms that control money and information. Harvard University Press, 2015. https://doi.org/10.4159/harvard.9780674736061

[24] Ziewitz, Malte. 'Governing algorithms: Myth, mess, and methods.' Science, Technology, & Human Values 41, no. 1 (2016): 3-16.

[25] Lawrence Lessig, 'Against Transparency,' The New Republic, Oct. 9, 2009 https://newrepublic.com/article/70097/against-transparency

[26] Michael Kearns, Seth Neel, Aaron Roth, Zhiwei Steven Wu, 'Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness,' Arxiv.org, Nov. 14, 2017 https://arxiv.org/abs/1711.05144

[27] Brent Mittelstadt, Sandra Wachter, 'Could Counterfactuals Explain Algorithmic Decisions Without Opening the Black Box?' Oxford Internet Instsitute blog, 15 Jan 2018 https://www.oii.ox.ac.uk/blog/could-counterfactuals-explain-algorithmic-decisions-without-opening-the-black-box/

[28] Nissenbaum, Helen. 'Computing and accountability.' Communications of the ACM 37, no. 1 (1994): 72-81.

[29] WENDELL WALLACH, A DANGEROUS MASTER 226-27 (2015); Betsy Cooper, Judges in Jeopardy!: Could IBM's Watson Beat Courts at Their Own Game?, 121 YALE L.J. ONLINE 87, 98 (2011)

[30] Louis Matsakis, 'Researchers fooled a Google AI into thinking a rifle was a helicopter', Wired, 20 Dec 2017 https://www.wired.com/story/researcher-fooled-a-google-ai-into-thinking-a-rifle-was-a-helicopter/

[31] Yvonne Baur, Brenda Reid, Steve Hunt, and Fawn Fitter, 'How AI Can End Bias,' Digitalist Magazine, 16 Jan 2017 https://www.digitalistmag.com/executive-research/how-ai-can-end-bias

[32] Tutt, Andrew, An FDA for Algorithms (March 15, 2016). 69 Admin. L. Rev. 83 (2017). Available at SSRN: https://ssrn.com/abstract=2747994 or http://dx.doi.org/10.2139/ssrn.2747994

[33] Barocas, Solon, and Andrew D. Selbst. 'Big data's disparate impact.' Cal. L. Rev. 104 (2016): 671.

[34] Ryan Calo, Robotics and the Lessons of Cyberlaw, 103 Cal. L. Rev. 513 (2015)

[35] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society 3(2). https://doi.org/10.1177/2053951716679679

[36] Forsyth, D. R. (2006). Conflict. In Forsyth, D. R. , Group Dynamics (5th Ed.) (P. 388 - 389) Belmont: CA, Wadsworth, Cengage Learning.

[37] Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. arXiv preprint arXiv:1609.07236.

[38] Hayek, F. (1978). New studies in philosophy, politics, economics, and the history of ideas. London: Routledge.

[39] Held, V. (1995). Justice and care: Essential readings in feminist ethics. Boulder, CO: Westview Press.

[40] Nussbaum, M. C. (1999). Sex and social justice. Oxford: Oxford University Press.

[41] Pérez-Garzón, C. A. (2018). Unveiling the meaning of social justice in Colombia. Mexican Law Review 10(2), 27–66.

[42] Rawls, J. (2001). Justice as fairness: A restatement. Cambridge, MA: The Belknap Press of Harvard University Press.

[43] Reisch, M. (2002). Defining social justice in a socially unjust world. Families in Society 83(4), 343–354.

[44] Novak, M. (2000). Defining social justice. First Things 108, 11-13.

[45] Rasinski, K. A. (1987). What's fair is fair--or is it? Value differences underlying public views about social justice. Journal of Personal and Social Psychology 53(1), 201–211.

[46] Michael Walzer, Spheres of Justice, (NY: Basic Books, 1983)

[47] Foster, A. D., and Rosenzweig, M. R. (2010). Microeconomics of technology adoption. Annual Review of Economics 2, 395–424.

[48] Gupta, A. (2018). AI in smart cities: Privacy, trust, and ethics. New Cities. Retrieved from: https://newcities.org/the-big-picture-ai-smart-cities-privacy-trust-ethics/

[49] Kamgar-Parsi-B., Lawson, W., and Kamgar-Parsi, B. (2011). Toward development of a face recognition system for watchlist surveillance. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(10), 1925-1937.

[50] Buolamwini, J., and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency: PMLR 81 (pp. 77-91).

[51] Simonite, T. (2018, January 11). When it comes to gorillas, Google Photos remains blind. WIRED. Retrieved from: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[52] Noble, S. (2018). Algorithms of oppression. New York: NYU Press.

[53] O'Neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2016.

[54] Algorithmic Justice League https://www.ajlunited.org/ Accessed on: 28 September 2018

[55] Puri, R. (2018, February 6). Mitigating bias in AI models. IBM.com. Retrieved from: https://www.ibm.com/blogs/research/2018/02/mitigating-bias-ai-models/

[56] Levin S. (2016). A beauty contest was judged by AI and the robots didn't like dark skin. The Guardian. Retrieved from https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

[57] Oswald, Marion, Jamie Grace, Sheena Urwin, and Geoffrey C. Barnes. 'Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental'proportionality.' Information & Communications Technology Law 27, no. 2 (2018): 223-250.

[58] Roettgers, J. (2018, February 2). 'Deepfakes' will create Hollywood's next sex tape scare. Variety. Retrieved from: https://variety.com/2018/digital/news/hollywood-sex-tapes-deepfakes-ai-1202685655/

[59] Romano, Aja. 'Jordan Peele's simulated Obama PSA is a double-edged warning against fake news.' Australasian Policing 10, no. 2 (2018): 44.

[60] iBeKabir. (2016, June 6). 'YOOOOOO LOOK AT THIS' [Twitter post]. Retrieved from: https://twitter.com/iBeKabir/status/740005897930452992

[61] HereroRocher. (2016, April 5). 'I saw a tweet saying 'Google unprofessional hairstyles for work'. I did. Then I checked the 'professional' ones' [Twitter post]. Retrieved from: https://twitter.com/HereroRocher/status/717457819864272896

[62] Cohn, E. (2015, October 4). Google image search has a gender bias problem. Huffington Post. Retrieved from: https://www.huffingtonpost.co.uk/entry/google-image-gender-bias_n_7036414

[63] Langston, J. (2015, April 9). Who's a CEO? Google image results can shift gender biases. University of Washington News. Retrieved from: http://www.washington.edu/news/2015/04/09/whos-a-ceo-google-image-results-can-shift-gender-biases/

[64] Vaidhynathan, S. (2012). The Googlization of everything. Berkeley: UC Berkeley Press.

[65] Koene, A., Perez, E., Webb, H., Patel, M., Ceppi, S., Jirotka, M., & McAuley, D. (2017). Editorial responsibilities arising from personalization algorithms. *ORBIT Journal*, *1*(1). https://doi.org/10.29297/orbit.v1i1.26

[66] The Internet of me: creating a personalized web experience. Swayy Shayna Hodkin. Wired. – https://www.wired.com/insights/2014/11/the-internet-of-me/ - Accessed 31/01/2017

[67] When the Internet Thinks It Knows You. Eli Pariser. The New York Times - http://www.nytimes.com/2011/05/23/opinion/23pariser.html - Accessed 31/01/2017.

[68] Should there be a better accounting of the algorithms that choose our news for us? David Sutcliffe. Oxford Internet Institute – https://www.oii.ox.ac.uk/should-there-be-a-better-accounting-of-the-algorithms-that-choose-our-news-for-us/ - Accessed 31/01/2017.

[69] The Trouble with the Echo Chamber Online. Natasha Singer. The New York Times – http://www.nytimes.com/2011/05/29/technology/29stream.html - Accessed 31/01/2017

[70] Colleone, E., Rossa, A., and Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. Journal of Communication 64(2), 317–332.

[71] Hong, S., and Kim, S. H. (2016). Political polarization on twitter: Implications for the use of social media in digital governments. Government Information Quarterly 33(4), 777–782.

[72] When algorithms discriminate. Claire Cain Miller. The Upshot. - https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html - Accessed 31/01/2017

[73] Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated experiments on ad privacy settings. arXiv preprint arXiv:1408.6491v2

[74] Sweeney, L. (2013). Discrimination in online ad delivery. Communications of the Association of Computing Machinery (CACM) 56(5), 44–54.

[75] boyd, D., and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication, and Society 15(5), 662–679.

[76] Dewey C. (2016) 98 personal data points that Facebook uses to target ads to you. The Washington Post. Retrieved from: https://www.washingtonpost.com/news/the-intersect/wp/2016/08/19/98-personal-data-points-that-facebook-uses-to-target-ads-to-you/

[77] Google Analytics Opt-out Browser Add-on. Accessed on: 28 September 2018 https://tools.google.com/dlpage/gaoptout

[78] DuckDuckGo 'The search engine that doesn't track you'. Accessed on: 28 September 2018 https://duckduckgo.com

[79] ProtonMail 'Get your encrypted email account'. Accessed on: 28 September 2018 https://protonmail.com/

[80] Andreou, Athanasios, Giridhari Venkatadri, Oana Goga, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 'Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations.' In The Network and Distributed System Security Symposium (NDSS). 2018. http://www.eurecom.fr/~andreou/papers/fb_ad_transparency_NDSS2018.pdf

[81] Meredith, S. (2018, April 10). Facebook-Cambridge Analytica: A timeline of the data hijacking scandal. CNBC. Retrieved from: https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html

[82] Lapowsky, I. (2018, April 4). Facebook exposed 87 million users to Cambridge Analytica. Wired. Retrieved from: https://www.wired.com/story/facebook-exposed-87-million-users-to-cambridge-analytica/

[83] Hogan, B. (2018). Social media giveth, social media taketh away: Facebook, friendships, and APIs. International Journal of Communication 12, 592–611.

[84] Chang, E (2018, June 8). The Former Facebook Exec Holding the Social Media Giant Accountable. Bloomberg Technology. Retrieved from: https://www.bloomberg.com/news/videos/2018-06-08/the-former-facebook-exec-holding-the-social-media-giant-accountable-video

[85] González, R. J. (2017). Hacking the citizenry?: Personality profiling, 'big data' and the election of Donald Trump. Anthropology Today 33(3), 9–12.

[86] Rosenberg, M., Confessore, N., and Cadwalladr, C. (2018, March 17). How Trump consultants exploited the Facebook data of millions. New York Times. Retrieved from: https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

[87] Confessore, N. (2018, April 4). Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. The New York Times. Retrieved from: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

[88] Depre, Joseph. 'May 14 May 14 YOU: Production and Projection and the End of Democracy.' https://www.hyper-psa.org/blog/2018/5/14/you-production-and-projection-and-the-end-of-democracy

[89] McClenaghan M. The 'Dark Ads' Election: How are political parties targeting you on facebook? The Bureau of Investigative Journalism. Retrieved from: https://www.thebureauinvestigates.com/stories/2017-05-15/the-dark-ads-election-how-are-political-parties-targeting-you-on-facebook#

[90] Solon, O. (2017, March 25). Google's bad week: YouTube loses millions as advertising row reaches US. The Guardian. Retrieved from: https://www.theguardian.com/technology/2017/mar/25/google-youtube-advertising-extremist-content-att-verizon

[91] Kehl, D., Guo, P., and Kessler, S. (2017). Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities. Retrieved from: https://cyber.harvard.edu/publications/2017/07/Algorithms

[92] Simourd, D. J. (2004). Use of dynamic risk/need assessment instruments among long-term incarcerated offenders. Criminal Justice and Behavior 31(3), 306–323.

[93] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. ProPublica. Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[94] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016, May 23). How we analyzed the COMPAS recidivism algorithm. Propublica. Retrieved from: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[95] Dieterich, W., Mendoza, C., Brennan, T. (2016). COMPAS risk scales: Accuracy equity, and predictive parity. Northpointe Inc. Research Department. Retrieved from: https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html

[96] Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Bilal Zafar, M. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. arXiv preprint arXiv:1807.00787.

[97] Spielkamp, M. (2017, June 12). Inspecting algorithms for bias. MIT Technology Review. Retrieved from: https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/

[98] Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 'Fairness through awareness.' In Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214-226. ACM, 2012.

[99] Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 'Fairness in criminal justice risk assessments: the state of the art.' arXiv preprint arXiv:1703.09207 (2017).

[100] Beriain, I. M. (2018). Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling. Law, Probability and Risk 17(1), 45-.–53.

[101] Durham Constabulary Police (2017), Written evidence submitted by Durham Constabulary (ALG0041), Algorithms in decision-making inquiry. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/written/69063.html

[102] Noack, R. (2016, May 28). How long would it take to read the terms of your smartphone apps? These Norwegians tried it out. The Washington Post. Retrieved from: https://www.washingtonpost.com/news/worldviews/wp/2016/05/28/how-long-would-it-take-to-read-the-terms-of-your-smartphone-apps-these-norwegians-tried-it-out/

[103] Yang, Ran, Chuang Lin, and Fujun Feng. 'A Time and Mutable Attribute-Based Access Control Model.' JCP 4, no. 6 (2009): 510-518.

[104] Meyer, D. (2018, May 25). AI has a big privacy problem and Europe's new data protection law Is about to expose it. Fortune. Retrieved from: http://fortune.com/2018/05/25/ai-machine-learning-privacy-gdpr/

[105] Cutts M. (2018, May 16) Why did our PageRank go down? Matt Cutts: Gadgets, Google, and SEO. Retrieved from: https://backlinko.com/google-ranking-factors

[106] Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2018). Beyond distributive fairness in algorithmic decision-making: Feature selection for procedurally fair learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA.

[107] Alexander, L. (2016, April 8). Do Google's 'unprofessional hair' results show it is racist? The Guardian. Retrieved from: https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist-

[108] Topping, A., and Belam, M. (2018, June 12). Campaign to change stereotypical search engine images of female football fans. The Guardian. Retrieved from: https://www.theguardian.com/football/2018/jun/12/campaign-to-change-stereotypical-search-engine-images-of-female-football-fans

[109] Guarino, B. (2016, June 10). Google faulted for racial bias in image search results for black teenagers. The Washington Post. Retrieved from: https://www.washingtonpost.com/news/morning-mix/wp/2016/06/10/google-faulted-for-racial-bias-in-image-search-results-for-black-teenagers/

[110] Webb, H. M., M. Patel, M. Rovatsos, A. Davoust, S. Ceppi, A. Koene, L. Dowthwaite, V. Portillo, M. Jirotka, and M. Cano. (2018). 'It would be pretty immoral to choose a random algorithm' Opening up algorithmic interpretability and transparency. ETHICOMP 2018

[111] Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. (2016). 'Inherent Trade-Offs in the Fair Determination of Risk Scores.' https://arxiv.org/pdf/1609.05807v1.pdf

[112] Chouldechova, Alexandra. (2017). 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments,' 1–17. https://doi.org/10.1089/big.2016.0047.

[113] Larson, J., and Angwin, J. (2016, July 29). Technical response to Northpointe. ProPublica. Retrieved from: https://www.propublica.org/article/technical-response-to-northpointe

[114] Nikhil Sonnad, 'Google Translate's gender bias pairs 'he' with 'hardworking' and 'she' with lazy, and other examples', Quartz, 29 Nov. 2017 https://qz.com/1141122/google-translates-gender-bias-pairs-he-with-hardworking-and-she-with-lazy-and-other-examples/

[115] Lee, D. (2016, March 25). Tay: Microsoft issues apology of racist chatbot fiasco. BBC News. Retrieved from: https://www.bbc.co.uk/news/technology-35902104

[116] Courtland, R. (2018, June 20). Bias detectives: the researchers striving to make algorithms fair. Nature. Retrieved from: https://www.nature.com/articles/d41586-018-05469-3

[117] Glauner, P., Valtchev, P., and State, R. (2018). Impact of biases in big data. arXiv preprint: https://arxiv.org/pdf/1803.00897.pdf

[118] Hautala, L. (2016). Google removes autocomplete suggestions about Jews, women. CNET. Retrieved from: https://www.cnet.com/news/google-removes-autocomplete-suggestions-about-jews-women/

[119] Miller, C. C. (2015, July 9). When algorithms discriminate. The New York Times. Retrieved from: https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html

[120] Goodman, B., and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a 'Right to Explanation.' AI Magazine 38(3). DOI: https://doi.org/10.1609/aimag.v38i3.2741

[121] Čuk, T., and van Waeyenberge, A. (2018). European legal framework for algorithmic and high frequency trading (Mifid 2 and MAR): A global approach to managing the risks of the modern trading paradigm. European Journal of Risk Regulation 9(1), 146–153.

[122] Bernard, Z. (2017, December 19). The first bill to examine 'algorithmic bias' in government agencies has just passed in New York City. Business Insider. Retrieved from: https://www.businessinsider.com/algorithmic-bias-accountability-bill-passes-in-new-york-city-2017-12

[123] Facebook AI Research, Accessed on: 28 September 2018 https://research.fb.com/category/facebook-ai-research/

[124] Artificial Intelligence at Google: Our Principles, Google AI, Accessed on: 28 September 2018 https://ai.google/principles/

[125] Wagner, B. (2018). Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping? In M. Hildebrandt (Ed.), Being Profiling. Cogitas ergo sum. Amsterdam: Amsterdam University Press.

[126] IEEE (2016), Ethically Aligned Design, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, Retrieved from: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf

[127] Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. Science and public policy 39(6), 751-760.

[128] Von Schomberg, R. (2013). A vision of responsible innovation. In R. Owen, M. Heintz and J Bessant (eds.), Responsible Innovation. London: John Wiley, forthcoming.

[129] Dressel, J., and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances 4(1). DOI: 10.1126/sciadv.aao5580

[130] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. (2017). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). ACM, New York, NY, USA, 275-284. DOI: https://doi.org/10.1145/3097983.3098066

[131] Mehta, S., Pimplikar, R., Singh, A., Varshney, L. R., and Visweswariah, K. (2013). Efficient multifaceted screening of job applicants. In Proceedings of the 16th International Conference on Extending Database Technology (pp. 661-671). New York: ACM.

[132] Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2016). Accountable algorithms. University of Pennsylvania Law Review 165, 633–706.

[133] U.S. Patent 20180032883: SOCIOECONOMIC GROUP CLASSIFICATION BASED ON USER FEATURES Retrieved from: http://appft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=20180032883&OS=20180032883&RS=20180032883

[134] Jackson, Peter. Introduction to expert systems. Addison-Wesley Longman Publishing Co., Inc., 1998.

[135] CRoss Industry Standard Process for Data Mining, CRISP-DM Consortium, 1999

[136] Azar, A.T. & El-Metwally, S.M. Neural Comput & Applic (2013) 23: 2387. https://doi.org/10.1007/s00521-012-1196-7

[137] V. Vapnik and A. Chervonenkis. Theory of Pattern Recognition. Nauka, Moscow, 1974

[138] Yishay Mansour. Pessimistic decision tree pruning. In Proceedings of Machine Learning, 195-201, 1997

[139] Koiran, Pascal, and Eduardo D. Sontag. 'Neural networks with quadratic VC dimension.' In Advances in neural information processing systems, pp. 197-203. 1996.

[140] Murat Kantarcioğlu, Jiashun Jin, and Chris Clifton. 2004. When do data mining results violate privacy?. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04). ACM, New York, NY, USA, 599-604. DOI=http://dx.doi.org/10.1145/1014052.1014126

[141] Dwork C., McSherry F., Nissim K., Smith A. (2006) Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi S., Rabin T. (eds) Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, vol 3876. Springer, Berlin, Heidelberg

[142] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2017. Guilt-free data reuse. Commun. ACM 60, 4 (March 2017), 86-93. DOI: https://doi.org/10.1145/3051088

[143] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2018. Never-ending learning. Commun. ACM 61, 5 (April 2018), 103-115. DOI: https://doi.org/10.1145/3191513

[144] Sullivan, Brendan M., Gopikrishna Karthikeyan, Zuli Liu, Wouter Lode Paul Massa, and Mahima Gupta. 'Socioeconomic group classification based on user features.' U.S. Patent Application 15/221,587, filed February 1, 2018.

[145] Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 'Accountability of AI under the law: The role of explanation.' arXiv preprint arXiv:1711.01134 (2017).

[146] Pfleeger, Shari Lawrence, and Joanne M. Atlee. Software engineering: theory and practice. Pearson Education India, 1998.

[147] Mäntylä, Mika V., and Casper Lassenius. 'What types of defects are really discovered in code reviews?.' IEEE Transactions on Software Engineering 35, no. 3 (2009): 430-448.

[148] Moritz Beller, Alberto Bacchelli, Andy Zaidman, and Elmar Juergens. 2014. Modern code reviews in open-source projects: which problems do they fix?. In Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014). ACM, New York, NY, USA, 202-211. DOI: http://dx.doi.org/10.1145/2597073.2597082

[149] Saltelli, Andrea, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. Sensitivity analysis in practice: a guide to assessing scientific models. John Wiley & Sons, 2004.

[150] Chen, Hongyi, and Dundar F. Kocaoglu. 'A sensitivity analysis algorithm for hierarchical decision models.' European Journal of Operational Research 185, no. 1 (2008): 266-288.

[151] Iman, Ronald L., and Jon C. Helton. 'An investigation of uncertainty and sensitivity analysis techniques for computer models.' Risk analysis 8, no. 1 (1988): 71-90.

[152] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar. 2011. Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence (AISec '11). ACM, New York, NY, USA, 43-58. DOI=http://dx.doi.org/10.1145/2046684.2046692

[153] Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 'Intriguing properties of neural networks.' arXiv preprint arXiv:1312.6199 (2013).

[154] Fawzi, A., Fawzi, O. & Frossard, P. Mach Learn (2018) 107: 481. https://doi.org/10.1007/s10994-017-5663-3

[155] Zhang, Junzhe, and Elias Bareinboim. 'Fairness in Decision-Making–The Causal Explanation Formula.' In 32nd AAAI Conference on Artificial Intelligence. 2018.

[156] American Association for the Advancement of Science. 'The Book of Why: The New Science of Cause and Effect.' (2018): 855-855.

[157] Pressman, Roger S. Software engineering: a practitioner's approach. Palgrave Macmillan, 2005.

[158] Fabian Benduhn, Thomas Thüm, Malte Lochau, Thomas Leich, and Gunter Saake. 2015. A Survey on Modeling Techniques for Formal Behavioral Verification of Software Product Lines. In Proceedings of the Ninth International Workshop on Variability Modelling of Software-intensive Systems (VaMoS '15). ACM, New York, NY, USA, , Pages 80 , 8 pages. DOI=http://dx.doi.org/10.1145/2701319.2701332

[159] Xavier Leroy. 2009. Formal verification of a realistic compiler. Commun. ACM 52, 7 (July 2009), 107-115. DOI: https://doi.org/10.1145/1538788.1538814

[160] Gerwin Klein, June Andronick, Kevin Elphinstone, Toby Murray, Thomas Sewell, Rafal Kolanski, and Gernot Heiser. 2014. Comprehensive formal verification of an OS microkernel. ACM Trans. Comput. Syst. 32, 1, Article 2 (February 2014), 70 pages. DOI=http://dx.doi.org/10.1145/2560537

[161] White, Neil, Stuart Matthews, and Roderick Chapman. 'Formal verification: will the seedling ever flower?.' Phil. Trans. R. Soc. A 375, no. 2104 (2017): 20150402.

[162] ISO/IEC JTC 1/SC 7 'Software and systems engineering', International Organization for Standardization (ISO), Retrieved from: https://www.iso.org/committee/45086.html

[163] Capability Maturity Model Integration, CMMI Institute, https://cmmiinstitute.com/

[164] IEEE P7001 Standards for Transparency of Autonomous Systems, IEEE-SA, https://standards.ieee.org/project/7001.html

[165] Winfield, Alan FT, and Marina Jirotka. 'The case for an ethical black box.' In Conference Towards Autonomous Robotic Systems, pp. 262-273. Springer, Cham, 2017.

[166] St Amant, Robert, Ralph Brewer, and MaryAnne Fields. Tracing Moral Agency in Robot Behavior. No. ARL-TN-0885. US Army Research Laboratory Aberdeen Proving Ground United States, 2018

[167] Fair Isaac Corporation FICO score, FICO, https://www1.myfico.com/products/onetimereports

[168] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 1135-1144. DOI: https://doi.org/10.1145/2939672.2939778

[169] Diakopoulos, Nicholas. 'Algorithmic-Accountability: the investigation of Black Boxes.' *Tow Center for Digital Journalism,* 2014. http://www.nickdiakopoulos.com/wp-content/uploads/2011/07/Algorithmic-Accountability-Reporting_final.pdf

[170] Mukherjee, Arjun, et al. 'What Yelp Fake Review Filter Might Be Doing?' Proceedings of the International Conference of Weblogs and Social Media (ICWSM), 2013.

[171] Shirriff, Ken. 'How Hacker News Ranking Really Works: Scoring, Controversy, and Penalties.' November 18, 2013. http://www.righto.com/2013/11/how-hacker-news-ranking-really-works.html

[172] Guha, Saikat, et al. 'Challenges in Measuring Online Advertising Systems.' Proc. Internet Measurement Conference (IMC), 2010.#

[173] Baker, Paul and Amanda Potts. ''Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms.' Critical Discourse Studies, 10 (2), 2013.

[174] 'Awesom-machine-learning-interpretability', list of open source tools for machine learning interpretability hosted on GitHub https://github.com/jphall663/awesome-machine-learning-interpretability

[175] IBM Research Trusted AI, 'AI Fairness 360 Open Source Toolkit', http://aif360.mybluemix.net/

[176] Chiandussi, G., Codegone, M., Ferrero, S. and Varesio, F.E., 2012. Comparison of multi-objective optimization methodologies for engineering applications. Computers & Mathematics with Applications, 63(5), pp.912-942.

[177] https://www.noesissolutions.com/technologies/design-space-exploration/multi-objective-optimization

[178] Mandal, J.K., Mukhopadhyay, S. and Dutta, P. eds., 2018. Multi-Objective Optimization: Evolutionary to Hybrid Framework. Springer.

[179] Francis, R. and Bekera, B., 2014. A metric and frameworks for resilience analysis of engineered and infrastructure systems. Reliability Engineering & System Safety, 121, pp.90-103.

[180] Pieters, W. 2011. Explanation and trust: what to tell the user in security and AI? Ethics Inf Technol 13: 53-64. https://doi.org/10.1007/s10676-010-9253-3

[181] Tufekci, Z., 2015. Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. J. on Telecomm. & High Tech. L., 13, p.203-216.

[182] Ethics Certifications Program for Autonomous and Intelligent Systems (ECPAIS) https://standards.ieee.org/industry-connections/ecpais.html

[183] Burgemeestre, Brigitte, Joris Hulstijn and Yao-Hua Tan. 'Rule-based versus Principle-based Regulatory Compliance.' JURIX (2009).

[184] R. B. Korobkin. Behavioral analysis and legal form: Rules vs. principles revisited. Oregon Law Review, 79(1):23 –60, 2000.

[185] Black, Julia, Forms and Paradoxes of Principles Based Regulation (September 23, 2008). LSE Legal Studies Working Paper No. 13/2008. Available at SSRN: https://ssrn.com/abstract=1267722 or http://dx.doi.org/10.2139/ssrn.1267722

[186] NAIC Response to Treas-DO-2007-0018 28th November 2007, available at http://www.naic.org/documents/topics_federal_regulator_treasury_response_0711.pdf.

[187] Ojo, M.2011. Building on the trust of management: Overcoming the paradoxes of principles based regulation. Banking & Financial Services Policy Report 30 (7): 1–9.

[188] C. L. Ford. Newgovernance, compliance, and principles-based securities regulation. American Business Law Journal, 45(1):1–60, 2008

[189] Schwarcz, S. L. 2009. The 'principles' paradox. European Business Organization Law Review 10: 175–184.

[190] L. A. Cunningham. A prescription to retire the rhetoric of principles-based systems in corporate law, securities regulation and accounting. Technical Report 127, Boston College Law School, 2007.

[191] Florian Saurwein, Natascha Just, Michael Latzer, (2015) 'Governance of algorithms: options and limitations', info, Vol. 17 Issue: 6, pp.35-49, doi: 10.1108/info-05-2015-0025 Permanent link to this document: http://dx.doi.org/10.1108/info-05-2015-0025

[192] Rothstein, Henry, Phil Irving, Terry Walden, and Roger Yearsley. 'The risks of risk-based regulation: Insights from the environmental policy domain.' Environment international 32, no. 8 (2006): 1056-1065.

[193] Posiva. The final disposal of spent nuclear fuel. Environmental Impact Assessment Report. General Summary. Helsinki, Finland: Posiva Oy. 1999. [available at http://www.posiva.fi/englanti/yv_ly.pdf]

[194] Knight, John C. 'Safety critical systems: challenges and directions.' In Proceedings of the 24th International Conference on Software Engineering, pp. 547-550. ACM, 2002.

[195] Wiad, Joseph. 'Software reuse: A safety-critical primer.' IEEE Aerospace and Electronic Systems Magazine 22, no. 4 (2007): 18-22.

[196] Jacklin, Stephen, Johann Schumann, Pramod Gupta, M. Lowry, John Bosworth, Eddie Zavala, Kelly Hayhurst, Celeste Belcastro, and Christine Belcastro. 'Verification, validation, and certification challenges for adaptive flight-critical control system software.' In AIAA Guidance, Navigation, and Control Conference and Exhibit, p. 5258. 2004.

[197] Mitka, Eleftheria, Antonios Gasteratos, Nikolaos Kyriakoulis, and Spyridon G. Mouroutsos. 'Safety certification requirements for domestic robots.' Safety science 50, no. 9 (2012): 1888-1897.

[198] Reed, Chris. 'Online and offline equivalence: Aspiration and achievement.' International journal of law and information technology 18, no. 3 (2010): 248-273.

[199] IBM Watson Health. Accessed on: 28 September 2018, https://www.ibm.com/watson/health/

[200] IBM Watson RegRach - Regulatory Technology for Banking and Financial Markets. Accessed on: 18 September 2018, https://www.ibm.com/industries/banking-financial-markets/risk-compliance

[201] Latzer, M., Just, N., Saurwein, F. and Slominski, P. (2002), Selbst- und Ko-Regulierung im Mediamatiksektor: Alternative Regulierungsformen zwischen Markt und Staat, Westdeutscher Verlag, Wiesbaden.

[202] Latzer, M., Just, N., Saurwein, F. and Slominski, P. (2003), 'Regulation remixed: institutional change through self and co-regulation in the mediamatics sector', Communications & Strategies, Vol. 50 No. 2, pp. 127-157

[203] Bartle, I. and Vass, P. (2005), Self-Regulation and the Regulatory State: A Survey of Policy and Practice, University of Bath School of Management.

[204] Nicholas Diakopoulos (2015) Algorithmic Accountability: Journalistic investigation of computational power structures, 3 Digital Journalism 3, 398-415, DOI: 10.1080/21670811.2014.976411

[205] Report By Aaron Rieke, Miranda Bogen, and David Robinson, 28 February 2018, Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods An Upturn and Omidyar Network Report http://omidyar.com/sites/default/files/file_archive/Public%20Scrutiny%20of%20Automated%20Decisions.pdf

[206] Katz, Michael L., and Carl Shapiro. 'Systems competition and network effects.' Journal of economic perspectives 8, no. 2 (1994): 93-115.

[207] Zittrain , J. 2008 . The future of the Internet and how to stop it . New Haven, CT : Yale University Press.

[208] Wu , T. 2010 . The master switch: The rise and fall of digital empires . New York : Knopf

[209] Brown, Ian, and Christopher T. Marsden. Regulating code: Good governance and better regulation in the information age. MIT Press, 2013.

[210] Breese, John S., David Heckerman, and Carl Kadie. 'Empirical analysis of predictive algorithms for collaborative filtering.' In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, pp. 43-52. Morgan Kaufmann Publishers Inc., 1998.

[211] Domingos, Pedro. 'A few useful things to know about machine learning.' Communications of the ACM 55, no. 10 (2012): 78-87.

[212] Awad, Naveen Farag, and Mayuram S. Krishnan. 'The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization.' MIS quarterly (2006): 13-28.

[213] Xu, Heng, Xin Robert Luo, John M. Carroll, and Mary Beth Rosson. 'The personalization privacy paradox: An exploratory study of decision-making process for location-aware marketing.' Decision support systems 51, no. 1 (2011): 42-52.

[214] Ebenezer, Mercy Elizabeth Devakirubai. 'The Impact of Consumer Privacy Behavior on the Purchase Decision Process of Smart Home Internet of Things (IoT) Devices.' PhD diss., 2017

[215] Hull, Gordon. 'Successful failure: what Foucault can teach us about privacy self-management in a world of Facebook and big data.' Ethics and Information Technology 17, no. 2 (2015): 89-101.

[216] Wilson, Dave, and Joseph S. Valacich. 'Unpacking the privacy paradox: Irrational decision-making within the privacy calculus.' (2012).

[217] Kollewe J. Marmite maker Unilever threatens to pull ads from Facebook and Google. The Guardian. Last Accessed on 28 September 2018 https://www.theguardian.com/media/2018/feb/12/marmite-unilever-ads-facebook-google

[218] Steurer, Reinhard. 'The role of governments in corporate social responsibility: Characterising public policies on CSR in Europe.' Policy sciences 43, no. 1 (2010): 49-72.

[219] How should I implement an Automated Decision System? (draft), Government of Canada Digital Playbook (draft), https://canada-ca.github.io/digital-playbook-guide-numerique/views-vues/automated-decision-automatise/en/automated-decision.html

[220] Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 'Algorithmic impact assessments: A practical framework for public agency accountability.' (2018). https://ainowinstitute.org/aiareport2018.pdf

[221] Mayor de Blasio Announces First-In-Nation Task Force To Examine Automated Decision Systems Used By The City, The Official Website of the City of New York, Accessed On: 28 September 2018. https://www1.nyc.gov/office-of-the-mayor/news/251-18/mayor-de-blasio-first-in-nation-task-force-examine-automated-decision-systems-used-by

[222] Schaar, P. (2010), 'Privacy by design', Identity in the Information Society, Vol. 3 No. 2, pp. 267-274.

[223] Cavoukia, A. (2012), 'Privacy by design: origins, meaning, and prospects for ensuring privacy and trust in the information era', available at: www.privacybydesign.ca/content/uploads/2010/03/PrivacybyDesignBook.pdf (accessed 12 August 2014).

[224] Munson, S.A. and Resnick, P. (2010), 'Presenting diverse political opinions: how and how much', Proceedings of ACM CHI 2010 Conference on Human Factors in Computing Systems 2010, Atlanta, Georgia, pp. 1457-1466.

[225] Schedl, M., Hauger, D. and Schnitzer, D. (2012), 'A model for serendipitous music retrieval', Proceedings of the 2nd Workshop on Context-awareness in Retrieval and Recommendation, Lisbon, pp. 10-13

[226] Resnick, P., Kelly Garrett, R., Kriplean, T., Munson, S.A. and Stroud, N.J. (2013), 'Bursting your (filter) bubble: strategies for promoting diverse exposure', Proceedings of the 2013 Conference on Computer-Supported Cooperative Work Companion, San Antonio, Texas, pp. 95-100.

[227] Chowdhury R., Tackling the challenge of ethics in AI, Digital Perspectives, Accenture Blog, Accessed on: 28 September 2018 https://www.accenture.com/gb-en/blogs/blogs-cogx-tackling-challenge-ethics-ai

[228] Power to the Tester. Accenture Testing Service. Accessed on: 28 September 2018 https://www.accenture.com/gb-en/service-application-testing-overview

[229] Matching Talent to Opportunity, Bias-Free, Pymetrics, Accessed on: 28 September 2018. https://www.pymetrics.com/employers/

[230] Shankland S. Facebook starts building AI with an ethical compass. CNET. Accessed on: 28 September 2018. https://www.cnet.com/news/facebook-starts-building-ai-with-an-ethical-compass/

[231] O'Neil Risk Consulting & Algorithmic Auditing, Accessed on: 28 September 2018. http://www.oneilrisk.com/

[232] ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*), Accessed on: 28 September 2018, http://fatconference.org/

[233] Artificial Intelligence, Ethics, and Society. AAAI/ACM conference on. Accessed on: 28 September 2018 http://www.aies-conference.com/

[234] Vallor S., An Ethical Toolkit for Engineering/Design Practice. Markkula Center for Applied Ethics, Santa Clara University, https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/

[235] Digital Decisions. Center for Democracy & Technology (CDT). Accessed on: 28 September 2018 https://cdt.org/issue/privacy-data/digital-decisions/

[236] Dialogues on AI and Ethics, Princeton University Center for Human Values, Accessed on: 28 September 2018. https://aiethics.princeton.edu/case-studies/

[237] DeepMind Ethics & Society, Accessed on: 20 September 2018, DeepMind Ethics & Society https://deepmind.com/applied/deepmind-ethics-society/

[238] Lin, P. and Selinger, E. (2014), 'Inside google's mysterious ethics board', Forbes, available at: www.forbes.com/sites/privacynotice/2014/02/03/inside-googles-mysterious-ethics-board/ (accessed on 29 September 2018)

[239] Bracha, O. and Pasquale, F. (2008), 'Federal search commission? Access, fairness and accountability in the law of search', Cornell Law Review, Vol. 93 No. 6, pp. 1149-1210.

[240] Granka, L.A. (2010), 'The politics of search: a decade retrospective', The Information Society, Vol. 26 No. 5, pp. 364-374.

[241] Rieder, B. (2005), 'Networked control: search engines and the symmetry of confidence', International Review of Information Ethics, Vol. 3, pp. 26-32.

[242] FTC – Federal Trade Commission (2014), 'Data brokers. A call for transparency and accountability', available at: www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency- accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf (accessed 12 August 2014).

[243] Anthes, Gary. 'Data brokers are watching you.' Communications of the ACM 58, no. 1 (2015): 28-30.

[244] Tiku N. Why Tech Workers Dissent is Going Viral. WIRED, Accessed on 28 September 2018 https://www.wired.com/story/why-tech-worker-dissent-is-going-viral/

[245] Murdock J., What is Project Maven? Google Urged to abandon U.S. Military Drone Program. Newsweek, Accessed on: 28 September 2018. https://www.newsweek.com/project-maven-google-urged-abandon-work-military-drone-program-926800

[246] Conger K. and Wakabayashi D., Google Employees Protest Secret Work on Censored Search Engine for China. The New York Times. Accessed on: 28 September 2018

[247] Simonite T., Google Sets Limits to its use of AI but allows defense work. WIRED, Accesses on 28 September 2018. https://www.wired.com/story/google-sets-limits-on-its-use-of-ai-but-allows-defense-work/

[248] Conger K., Google Plans Not to Renew Its Contract for Project Mave, a Controversial Pentagon Drone AI imaging Program. Accessed on: 28 September 2018.https://gizmodo.com/google-plans-not-to-renew-its-contract-for-project-mave-1826488620

[249] Former Google and Facebook Staff Launch Anti-Addiction Lobby. RED Herring, Accessed on: https://www.redherring.com/consumer-electronics/former-google-facebook-staff-launch-anti-addiction-lobby/

[250] Evangelista B., Check your phone 86 times a day? Tech insiders say that's by design. San Francisco Chronicle. Accessed on 28 September 2018. https://www.sfchronicle.com/business/article/Check-your-phone-86-times-a-day-Tech-insiders-12888512.php

[251] Wilson M., Google's Plan to make tech less addictive, Fastcompany, Accessed on: 28 September 2018. https://www.fastcompany.com/90171307/googles-plan-to-make-tech-less-addictive

[252] Weber H., Apple is Trying to Make Your iPhone Less Addictive, GIZMODO, Accessed on: 28 September 2018 https://gizmodo.com/apple-is-trying-to-make-ios-12-less-addictive-with-new-1826530852

[253] Sandoval G., Over 100 Amazon employees, including senior software engineers, signed a letter asking Jeff Bezos to stop selling facial recognition software to police. Accesses on: 28 September 2018. http://uk.businessinsider.com/over-100-amazon-employees-sign-letter-jeff-bezos-stop-selling-facial-recognition-software-police-2018-6

[254] Wexler, James. 'The What-If Tool: Code-Free Probing of Machine Learning Models'. Google AI Blog. 11 Sept. 2018. https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

[255] AI Fairness 360 Open Source Toolkit, IBM, Accessed on: 28 September 2018. https://aif360.mybluemix.net/

[256] Teich, Paul. 'Artificial Intelligence Can Reinforce Bias, Cloud Giants Announce Tools For AI Fairness.' Forbes, 24 Sept. 2018. https://www.forbes.com/sites/paulteich/2018/09/24/artificial-intelligence-can-reinforce-bias-cloud-giants-announce-tools-for-ai-fairness/#362f66cc9d21

[257] Anderson, Ronald E. 'ACM code of ethics and professional conduct.' Communications of the ACM 35, no. 5 (1992): 94-99.

[258] Chatila, Raja, Kay Firth-Butterfield, John C. Havens, and Konstantinos Karachalios. 'The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems [standards].' IEEE Robotics & Automation Magazine 24, no. 1 (2017): 110-110.

[259] Schwab K., This logo is like an 'organic' sticker for algorithms. Fastcompany, Accessed on: 28 September 2018

[260] Hern, Alex. ''Partnership on AI' formed by Google, Facebook, Amazon, IBM and Microsoft.' The Guardian 28 (2016). Last Accessed: https://www.partnershiponai.org/

[261] ACM Code of Ethic and Professional Conduct. ACM. Accessed on 28 September 2018, https://www.acm.org/code-of-ethics

[262] IEEE Code of Conduct, IEEE. Accessed on 28 September 2018. https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/ieee_code_of_conduct.pdf

[263] Future of Life Institute, 2017. Asilomar AI Principles. Future of Life Institute, [online] Available at: https://futureoflife.org/ai-principles/

[264] Fuerguson, S., Thornley, C. and Gibb, F., 2016. Beyond codes of ethics. International Journal of Information Management: The Journal for Information Professionals, [e-journal] 36(4) pp.543-556. Available at: https://dl.acm.org/citation.cfm?id=2940433

[265] Consultancy.uk., 2018. The top five ethical moral principles for digital transformation. Consultancy.uk, [blog] 11 April. Accessed on: https://www.consultancy.uk/news/16602/the-top-five-ethical-moral-principles-for-digital-transformation

[266] Mind of the Universe - Robots in Society: Blessing or Curse?, TU Delft, Accessed on: 28 September 2018. https://www.edx.org/course/mind-of-the-universe-robots-in-society-blessing-or-curse

[267] Zittrain J. and Ito J., The Ethics and Governance of Artificial Intelligence, MIT Media Lab, Accessed on: 28 September 2018. https://www.media.mit.edu/courses/the-ethics-and-governance-of-artificial-intelligence/

[268] CS122: Artificial Intelligence - Philosophy, Ethics, and Impact. Stanford University. Accessed on: 28 September 2018. http://web.stanford.edu/class/cs122/

[269] Baron, Justus, and Daniel F. Spulber. 'Technology standards and standard setting organizations: Introduction to the searle center database.' Journal of Economics & Management Strategy 27, no. 3 (2018): 462-503.

[270] Attia, John O., Dhadesugoor Vaman, and Matthew NO Sadiku. 'Engineering Standards: An Introduction for Electrical and Computer Engineering Students.' European Scientific Journal, ESJ 13, no. 9 (2017). https://eujournal.org/index.php/esj/article/download/9028/8613

[271] W3C Standards, W3C. Accessed on: 28 September 2018. https://www.w3.org/standards/

[272] ISO/IEC/IEEE 29119 Software Testing. The International Software Testing Standard. Accessed on: 28 September 2018. http://softwaretestingstandard.org/

[273] IEEE 1012-2016 - IEEE Standard for System, Software, and Hardware Verification and Validation. IEEE Standards Association. Accessed on: 28 September 2018. https://standards.ieee.org/standard/1012-2016.html

[274] ISO/IEC 27000 family - Information security management systems, International Standards Organization (ISO). Accessed on: 28 September 2018. https://www.iso.org/isoiec-27001-information-security.html

[275] Cybersecurity Standards. IT Governance. Accessed on: 28 September 2018. https://www.itgovernance.co.uk/cybersecurity-standards

[276] ISO/IEC/IEEE 29148:2011 Systems and software engineering -- Life cycle processes -- Requirements engineering. International Standards Organization (ISO). Accessed on: 28 September 2018. https://www.iso.org/standard/45171.html

[277] IEEE 1063-2001 - IEEE Standard for Software User Documentation. IEEE Standards Association. Accessed on: 28 September 2018. https://standards.ieee.org/standard/1063-2001.html

[278] 14764-2006 - ISO/IEC/IEEE International Standard for Software Engineering - Software Life Cycle Processes - Maintenance. IEEE Standards Association. Accessed on: 28 September 2018. https://ieeexplore.ieee.org/document/1703974

[279] IEEE Computer Society, IEEE Std 1028 - IEEE Standard for Software Reviews and Audits. IEEE Standards Association. Accessed on: 28 September 2018. http://ieeexplore.ieee.org/document/4601584

[280] IT Standards. IT Governance. Accessed on: 28 September 2018. https://www.itgovernance.co.uk/standards

[281] IEEE P7003 Standards for Algorithmic Bias Considerations, IEEE-SA, Accessed on: 28 September 2018. https://standards.ieee.org/project/7003.html

[282] ISO/IEC JTC 1/SC 42 Artificial Intelligence. International Standards Organization (ISO). Accessed on: 28 September 2018. https://www.iso.org/committee/6794475.html

[283] Forsstrom, J. (1997). Why certification of medical software would be useful? Medical Informatics Research Centre, University of Turku, Turku, December 1997; 47(3) pp. 143–152.

[284] Heck, P., Klabbers, M. & van Eekelen, M. Software Qual J (2010) 18: 37. https://doi.org/10.1007/s11219-009-9080-0

[285] Ferreira, Gabriel. 'Software certification in practice: how are standards being applied?.' In Software Engineering Companion (ICSE-C), 2017 IEEE/ACM 39th International Conference on, pp. 100-102. IEEE, 2017.

[286] Common Criteria for Information Technology Security Evaluation, Version 3.1 Revision 4. [Online]. Available: http://www.commoncriteriaportal.org/cc/.

[287] RTCA. 1992. DO-178C–Software Considerations in Airborne Systems and Equipment Certification. Radio and Technical Commission for Aeronautics

[288] Anisetti, Marco, Claudio Ardagna, Ernesto Damiani, and Filippo Gaudenzi. 'A semi-automatic and trustworthy scheme for continuous cloud service certification.' IEEE Transactions on Services Computing (2017).

[289] Stephanow, Philipp, and Koosha Khajehmoogahi. 'Towards Continuous Security Certification of Software-as-a-Service Applications Using Web Application Testing Techniques.' In Advanced Information Networking and Applications (AINA), 2017 IEEE 31st International Conference on, pp. 931-938. IEEE, 2017.

[290] Edwards, Lilian, and Michael Veale. 'Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?.' IEEE Security & Privacy 16, no. 3 (2018): 46-54.

[291] D.K. Citron, 'Technological Due Process,' Washington University Law Review, vol. 85, 2008, pp. 1249–1313.

[292] K. Crawford and J. Schultz, 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms,' Boston College Law Review, vol. 55, 2014, pp. 93–128.

[293] Hunter, Philip. 'The big health data sale: As the trade of personal health and medical data expands, it becomes necessary to improve legal frameworks for protecting patient anonymity, handling consent and ensuring the quality of data.' EMBO reports (2016): e201642917.

[294] Neff, Gina, and Peter Nagy. 'Automation, algorithms, and politics| talking to bots: symbiotic agency and the case of tay.' International Journal of Communication 10 (2016): 17.

[295] Kramer, Adam DI, Jamie E. Guillory, and Jeffrey T. Hancock. 'Experimental evidence of massive-scale emotional contagion through social networks.' Proceedings of the National Academy of Sciences (2014): 201320040.

[296] Partnership on AI. Accessen on: 28 September 2018. https://www.partnershiponai.org/

[297] Shead S. The biggest mystery in AI right now is the ethics board that Google set up after buying DeepMind. Business Insider UK. Accessed on: 28 September 2018. http://uk.businessinsider.com/google-ai-ethics-board-remains-a-mystery-2016-3

[298] Hern A. Whatever happened to the DeepMind AI ethics board Google promised? The Guardian. Accessed on: 28 September 2018. https://www.theguardian.com/technology/2017/jan/26/google-deepmind-ai-ethics-board

[299] Meet the Partners. Partnership on AI. Accessed on: 28 September 2018. https://www.partnershiponai.org/partners/

[300] Hirsch, Dennis D. 'The law and policy of online privacy: Regulation, self-regulation, or co-regulation.' Seattle UL Rev. 34 (2010): 439.

[301] Guidelines for Online Privacy Policies, ONLINE PRIVACY ALLIANCE, http://www.privacyalliance.org/resources/ppguidelines/ [last visited 23 Sept 2018]

[302] Lee, Edward. 'Recognizing Rights in Real Time: The Role of Google in the EU Right to Be Forgotten.' UCDL Rev. 49 (2015): 1017. http://scholarship.kentlaw.iit.edu/cgi/viewcontent.cgi?article=3407&context=fac_schol

[303] Albareda, Laura. 'Corporate responsibility, governance and accountability: from self-regulation to co-regulation.' Corporate Governance: The international journal of business in society 8, no. 4 (2008): 430-439.

[304] Mandelkern Group on Better Regulation, Final Report, (November 13, 2001) available at http://europa.eu.int/comm/secretariat_general/impact/docs/mandelkern.pdf

[305] Hanson, David. CE marking, product standards and world trade. Edward Elgar Publishing, 2005.

[306] CE marking. Your Europe, European Union. Accessed on: 28 September 2018. https://europa.eu/youreurope/business/product/ce-mark/index_en.htm

[307] Felini, Damiano. 'Beyond today's video game rating systems: A critical approach to PEGI and ESRB, and proposed improvements.' Games and Culture 10, no. 1 (2015): 106-122.

[308] Brand, Jeffrey. 'A comparative analysis of ratings, classification and censorship in selected countries around the world.' (2002): 1.

[309] Weiss, Martin A., and Kristin Archick. 'US-EU data privacy: from safe harbor to privacy shield.' (2016).

[310] Tracol, Xavier. 'EU–US Privacy shield: the saga continues.' Computer Law & Security Review 32, no. 5 (2016): 775-777.

[311] European Governance – A White Paper, COM (2001) 428 final, (July 25, 2001), available at http://europa.eu.int/eur- lex/en/com/cnc/2001/com2001_0428en01.pdf

[312] Environmental Impact Assessment -EIA. European Commission - Environment. Accessed on: 28 September 2018. http://ec.europa.eu/environment/eia/eia-legalcontext.htm

[313] Glasson, John, and Riki Therivel. Introduction to environmental impact assessment. Routledge, 2013.

[314] Brown, Ian, and Christopher T. Marsden. Regulating code: Good governance and better regulation in the information age. MIT Press, 2013.

[315] Senden , L. 2005 . Soft law, self-regulation and co-regulation in European law: Where do they meet? Electronic Journal of Comparative Law 9 (1) 1 – 27 .

[316] Hüpkes , Eva . 2009 . Regulation, self-regulation or co-regulation? Journal of Business Law 5 : 427 – 446.

[317] Lievens, Eva, Jos Dumortier, and Patrick S. Ryan. 'The co-protection of minors in new media: A European approach to co-regulation.' UC Davis J. Juv. L. & Pol'y 10 (2006): 97.

[318] Frydman , B. , L. Hennebel , and G. Lewkowicz . 2008 . Public strategies for Internet co-regulation in the United States, Europe and China. http://papers.ssrn.com/sol3/ papers.cfm?abstract_id=1282826

[319] Tarlach McGonagle, Practical and Regulatory Issues Facing the Media Online, in SPREADING THE WORD ON THE INTERNET: 16 ANSWERS TO 4 QUESTIONS 94 (Christiane Hardy & Christian Möller eds. 2003)

[320] Neil Gunningham & Darren Sinclair, Leaders And Laggards: Next-Generation Environmental Regulation 104–05

[321] Lyle Scruggs, Sustaining Abundance: Environmental Performance In Industrial Democracies 145–46 (2003)

[322] Joseph Rees, Reforming The Workplace: A Study Of Self-Regulation In Occupational Safety (1988).

[323] Bert-Jaap Koops, Miriam Lips, Sjaak Nouwt, Corien Prins & Maurice Schellekens, Should Self-Regulation be the Starting Point?, in STARTING POINT FOR ITC REGULATION: DECONSTRUCTING PREVALENT POLICY ONE-LINERS 109, 149 (Bert-Jaap Koops, Corien Prins, Maurice Schellekens & Miriam Lips eds., 2006)

[324] Colin J. Bennett & Charles D. Raab, The Governance Of Privacy: Policy Instruments In Global Perspective 134 (2003)

[325] Lemley, Mark A., and David McGowan. 'Legal implications of network economic effects.' Calif. L. Rev. 86 (1998): 479.

[326] Varian, Hal R. 'High-technology industries and market structure.' University of California, Berkeley 33 (2001).

[327] Browne, Glenn J., and Nirup M. Menon. 'Network effects and social dilemmas in technology industries.' IEEE software 21, no. 5 (2004): 44-50.

[328] People, Power and Technology: The 2018 Digital Understanding Report, Doteveryone, Accessed on: 28 Spetember 2018. http://understanding.doteveryone.org.uk/files/Doteveryone_PeoplePowerTechDigitalUnderstanding2018.pdf

[329] Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 'Fair, Transparent, and Accountable Algorithmic Decision-making Processes.' Philosophy & Technology (2017): 1-17.

[330] Klawitter, Erin, and Eszter Hargittai. ''It's Like Learning a Whole Other Language:' The Role of Algorithmic Skills in the Curation of Creative Goods.' International Journal of Communication 12 (2018): 21.

[331] boyd, d., Crawford, K.: Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication, & Society 15(5), 662{679 (2012)

[332] Bhargava, R., Deahl, E., Letouze, E., Noonan, A., Sangokoya, D., Shoup, N.: Beyond data literacy: Reinventing community engagement and empowerment in the age of data. Data-Pop Alliance White Paper Series (2015). URL http://datapopalliance.org/wp-content/uploads/2015/11/Beyond-Data- Literacy-2015.pdf

[333] Burrell, J. (2016). How the machine `thinks': Understanding opacity in machine learning algorithms. Big Data & Society 3(1)

[334] Elvira Perez Vallejos, Ansgar Koene, Virginia Portillo, Liz Dowthwaite and Monica Cano, 'Young people's policy recommendations on algorithm fairness', in Proceedings of the 2017 ACM on Web Science Conference, Pages 247-251 ISBN: 978-1-4503-4896-6 DOI: 10.1145/3091478.3091512

[335] Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., . . . Sandvig, C. (2015). '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 (pp. 153–162). New York, NY: Association for Computing Machinery.

[336] Papsdorf, Christian, and Sebastian Jakob. 'Ein Kampf gegen Windmühlen: Jugendliche und junge Erwachsene im Umgang mit Algorithmen und Überwachung im Internet.' kommunikation@ gesellschaft 18 (2017): 27.

[337] Moses, Lyria Bennett. 'Is Your Algorithm Dangerous?[Leading Edge].' IEEE Technology and Society Magazine 37, no. 3 (2018): 20-21.

[338] Baker, Jamie J. 'Beyond the Information Age: The Duty of Technology Competence in the Algorithmic Society.' SCL Rev. 69 (2017): 557.

[339] Buchanan, B., and Taylor, M. 2017. 'Machine Learning for Policymakers,' Paper, Cyber Security Project, Belfer Center. https://www.belfercenter.org/publication/machine-learning-policymakers

[340] Oliver, Nuria. '22 The tyranny of data? The bright and dark sides of algorithmic decision-making for public policy making.' Assessing the impact of machine intelligence on human behaviour: an interdisciplinary endeavour (2018): 58.

[341] Pasquale, F.: The Black Blox Society: The secret algorithms that control money and information. Harvard University Press (2015)

[342] Rainie, Lee, and Anderson, Janna. 2017. 'Theme 7: The Need Grows for Algorithmic Literacy, Transparency and Oversight.' http://www.pewinternet.org/2017/02/08/theme-7-the-need-grows-for-algorithmic-literacy-transparency-and-oversight/.

[343] Kaminski, Margot. 2018. 'The Right to Explanation, Explained.' LawArXiv. June 19. doi:10.31228/osf.io/rgeus.

[344] Selbst, Andrew D., and Julia Powles. 'Meaningful information and the right to explanation.' International Data Privacy Law 7, no. 4 (2017): 233-242.

[345] Ananny, Mike, and Kate Crawford. 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability.' New Media & Society 20, no. 3 (2018): 973-989.

[346] Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 'The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards.' arXiv preprint arXiv:1805.03677 (2018).

[347] Yang, Ke, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. 'A Nutritional Label for Rankings.' In Proceedings of the 2018 International Conference on Management of Data, pp. 1773-1776. ACM, 2018.

[348] Perel, Maayan, and Niva Elkin-Koren. 'Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement.' Fla. L. Rev. 69 (2017): 181.

[349] Ananny, Mike, and Kate Crawford. 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability.' New Media & Society 20, no. 3 (2018): 973-989.

[350] Diakopoulos, Nicholas. 'Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens.' In Transparent Data Mining for Big and Small Data, pp. 25-43. Springer, Cham, 2017.

[351] Tesfay, Welderufael B., Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 'I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR.' In Companion of the The Web Conference 2018 on The Web Conference 2018, pp. 163-166. International World Wide Web Conferences Steering Committee, 2018.

[352] Jones, Rhianne, Neelima Sailaja, and Lianne Kerlin. 'Probing the design space of usable privacy policies: a qualitative exploration of a reimagined privacy policy.' In Proceedings of the 31st British Computer Society Human Computer Interaction Conference, p. 50. BCS Learning & Development Ltd., 2017.

[353] Cranor, L. F., Hoke, C., Leon, P. G., & Au, A. (2014). Are They Worth Reading? An In-Depth Analysis of Online Advertising Companies' Privacy Policies. Presented at the 42nd Research Conference on Communication, Information and Internet Policy. Available at http://www.contrib.andrew.cmu.edu/~pgl/tprc2014.pdf

[354] Rader, Emilee, Kelley Cotter, and Janghee Cho. 'Explanations as Mechanisms for Supporting Algorithmic Transparency.' In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, p. 103. ACM, 2018.

[355] John O. McGinnis, Accelerating AI, 104 NW. U. L. REV. 1253, 1262 (2010).

[356] Data Transparency Lab, Accessed on: 28 September 2018. http://www.datatransparencylab.org/

[357] Gunning, D., Explainable Artificial Intelligence (XAI). DARPA. Accessed on: 28 September 2018. https://www.darpa.mil/program/explainable-artificial-intelligence

[358] International Neuroinformatics Coordinating Facility (INCF). Accessed on: 28 September 2018. https://www.incf.org/

[359] Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe et al. 'The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.' arXiv preprint arXiv:1802.07228 (2018).

[360] Cambridge Analytica. The Guardian. Accessed on: 28 September 2018. https://www.theguardian.com/uk-news/cambridge-analytica

[361] Winston A., Palantir has secretly been using new orleans to test its predictive policing technology. The Verge. Accesses on: 28 September 2018. https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd

[362] Lecher C., What happens when an algorithm cuts your health care. The Verge. Accessed on: 28 September 2018. https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy

[363] OpenSCHUFA - shedding light on Germany's opaque credit scoring. AlgorithmWatch. Accessed on: 28 September 2018. https://algorithmwatch.org/en/openschufa-shedding-light-on-germanys-opaque-credit-scoring/

[364] Howard, Alexander Benjamin. 'The art and science of data-driven journalism.' (2014). https://doi.org/10.7916/D8Q531V1

[365] Arthur, W. Brian. 'Competing technologies, increasing returns, and lock-in by historical events.' The economic journal 99, no. 394 (1989): 116-131.

[366] Foxon, Timothy J. 'Technological lock-in and the role of innovation.' Handbook of sustainable development (2007): 140-152.

[367] Jaffe, Adam B., and Karen Palmer. 'Environmental regulation and innovation: a panel data study.' Review of economics and statistics 79, no. 4 (1997): 610-619.

[368] Horbach, Jens, Christian Rammer, and Klaus Rennings. 'Determinants of eco-innovations by type of environmental impact—The role of regulatory push/pull, technology push and market pull.' Ecological economics 78 (2012): 112-122.

[369] van Lieshout, Marc, and Sophie Emmert. 'RESPECT4U–privacy as innovation opportunity.'

[370] Goodman, Bryce, and Seth Flaxman. 'European Union regulations on algorithmic decision-making and a' right to explanation'.' arXiv preprint arXiv:1606.08813 (2016).

[371] Gulbenkoglu E., How to Build Explainable AI. DataEthics. Accessed on: 28 September 2018. https://dataethics.eu/en/how-to-built-explainable-ai/

[372] Kotonya, Gerald, and Ian Sommerville. 'Requirements Engineering: Processes and Techniques. 1998.' (1998).

[373] Glinz, Martin. 'On non-functional requirements.' In Requirements Engineering Conference, 2007. RE'07. 15th IEEE International, pp. 21-26. IEEE, 2007.

[374] Chung, Lawrence, and Julio Cesar Sampaio do Prado Leite. 'On non-functional requirements in software engineering.' In Conceptual modeling: Foundations and applications, pp. 363-379. Springer, Berlin, Heidelberg, 2009.

[375] Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 'Why a right to explanation of automated decision-making does not exist in the general data protection regulation.' International Data Privacy Law 7, no. 2 (2017): 76-99.

[376] Edwards, Lilian, and Michael Veale. 'Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking for.' Duke L. & Tech. Rev. 16 (2017): 18.

[377] Vedder, Anton, and Laurens Naudts. 'Accountability for the use of algorithms in a big data environment.' International Review of Law, Computers & Technology 31, no. 2 (2017): 206-224.

[378] LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique, LA CIRCULATION DES DONNÉES ET DU SAVOIR. Legifrance. Accessed on: 28 September 2018. https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo

[379] M.T. Ribeiro et al., 'Why Should I Trust You?: Explaining the Predictions of Any Classifier,' Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[380] G. Montavon et al., 'Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition,' Pattern Recognit., vol. 65, 2017, pp. 211–222.

[381] Ryan Poplin, Avinash V. Varadarajan, Katy Blumer, Yun Liu, Michael V. McConnell, Greg S. Corrado, Lily Peng, Dale R. Webster, 'Predicting Cardiovascular Risk Factors from Retinal Fundus Photographs using Deep Learning', NAture Biomedical Engineering, vol. 2. March 2018, pp. 158–164. Preprint available here: https://arxiv.org/pdf/1708.09843.pdf

[382] Weinberger, David. 'Optimization over Explanation: Maximizing the benefits of machine learning without sacrificing its intelligence.' 17 Jan. 2018 https://medium.com/berkman-klein-center/optimization-over-explanation-41ecb135763d

[383] Čerka, Paulius, Jurgita Grigienė, and Gintarė Sirbikytė. 'Liability for damages caused by artificial intelligence.' Computer Law & Security Review 31, no. 3 (2015): 376-389.

[384] Schmehl, Ian, and Stephen E. Lunce. 'Culpabilities in Medical Diagnostic Software: A Review of Legal Implications.' AMCIS 2000 Proceedings (2000): 103.

[385] Price, W. Nicholson, Medical Malpractice and Black-Box Medicine (February 2, 2017). I. Glenn Cohen et al., eds., Big Data, Health Law, and Bioethics (Cambridge University Press, 2018); U of Michigan Public Law Research Paper No. 536. Available at SSRN: https://ssrn.com/abstract=2910417

[386] Ryan Calo, Robotics and the Lessons of Cyberlaw, 103 Cal. L. Rev. 513 (2015)

[387] Cass R. Sunstein, The Ethics of Nudging, 32 YALE J. ON REG. 413, 414 (2015)

[388] Eric A. Posner, E. Glen Weyl, An FDA for Financial Innovation: Applying the Insurable Interest Doctrine to Twenty-First-Century Financial Markets, 107 NW. U. L. REV. 1307, 1355 (2013)

[389] Wachter, Sandra and Mittelstadt, Brent, A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI (September 13, 2018). Columbia Business Law Review, forthcoming (2019). Available at SSRN: https://ssrn.com/abstract=3248829

[390] Brent Daniel Mittelstadt and Luciano Floridi, ‘The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts’ (2016) 22 Science and Engineering Ethics 303.

[391] Paul Ohm, 'The Fourth Amendment in a World without Privacy' (2011) 81 Miss. LJ 1309; Pauline T Kim, 'Data-Driven Discrimination at Work' 58 81.

[392] Mittelstadt (n 27); Luciano Floridi, 'The Informational Nature of Personal Identity' (2011) 21 Minds and Machines 549.

[393] Sandra Wachter, 'Privacy: Primus Inter Pares — Privacy as a Precondition for Self-Development, Personal Fulfilment and the Free Enjoyment of Fundamental Human Rights' (Social Science Research Network 2017) SSRN Scholarly Paper ID 2903514. Accessed on: 28 September 2018. https://papers.ssrn.com/abstract=2903514

[394] Urteil des Ersten Senats vom BVerfG, '15. Dezember 1983, 1 BvR 209/83: Volkszählungsurteil' URL: http://dejure. org/dienste/vernetzung/rechtsprechung. Judgement of German Constitutional Court, BVerfG · Urteil vom 15. Dezember 1983 · Az. 1 BvR 209/83, 1 BvR 484/83, 1 BvR 420/83, 1 BvR 362/83, 1 BvR 269/83, 1 BvR 440/83 (Volkszählungsurteil).

[395] Omer Tene and Jules Polonetsky, 'Big Data for All: Privacy and User Control in the Age of Analytics' (2012) 11 Nw. J. Tech. & Intell. Prop. xxvii, 270.

[396] Joris van Hoboken, Search Engine Freedom: On the Implications of the Right to Freedom of Expression for the Legal Governance of Web Search Engines (Kluwer Law International Den Haag 2012);

[397] Joris van Hoboken, 'The Proposed Right to Be Forgotten Seen from the Perspective of Our Right to Remember' [2013] Freedom of Expression Safeguards in a Converging Information Environment, Prepared for the European Commission, Amsterdam.

[398] Larsson, Stefan. 'Algorithmic governance and the need for consumer empowerment in data-driven markets.' Internet Policy Review 7, no. 2 (2018).

[399] de Streel, Alexandre, and Anne-Lise Sibony. 'Towards Smarter Consumer Protection Rules for Digital Services.' (2017).

[400] Larsson, S. (2017a). All-seeing giants and blindfolded dwarfs: On information-asymmetries on data-driven markets. In J. Lith (Ed.), New Economic Models: Tools for Political Decision Makers Dealing with the Changing European Economies. Brussels, Belgium: European Liberal Forum asbl. Available at http://lup.lub.lu.se/record/bada07c0-3a62-4e12-950d-779178eeccd4

[401] Larsson, S. (2017c). Sustaining Legitimacy and Trust in a Data-driven Society. Ericsson Technology Review, 94(1), 40-49. Available at https://lup.lub.lu.se/search/publication/75b9d975-1a58-4145-85c4-efde2e46aa14

[402] Rhoen, M. (2016). Beyond consent: improving data protection through consumer protection law. Internet Policy Review, 5(1). doi:10.14763/2016.1.404

[403] Pasquale, F. (2017, September 12). Exploring the Fintech Landscape. Written Testimony of Frank Pasquale Before the United States Senate Committee on the Banking, Housing, and Urban Affairs. Available at https://www.banking.senate.gov/imo/media/doc/Pasquale%20Testimony%209-12-17.pdf

[404] European Data Protection Supervisor. (2015). Meeting the challenges of big data: A call for transparency, user control, data protection and accountability (Opinion No. 7/2015). Brussels: European Data Protection Supervisor. Available at https://edps.europa.eu/data-protection/our-work/publications/opinions/meeting-challenges-big-data_en

[405] King, N.J. & Forder, J. (2016). Data analytics and consumer profiling: Finding appropriate privacy principles for discovered data, Computer Law & Security Review, 32(5), 696-714. doi:10.1016/j.clsr.2016.05.002

[406] Naveen Kashyap, Why Pfizer Won in the United States but Lost in Canada, and the Challenges of Pharmaceutical Industry, 16 T.M. COOLEY J. PRAC. & CLINICAL L. 189, 202-03 (2014).

[407] John N. Joseph et. al., Enforcement Related to Off-Label Marketing and Use of Drugs and Devices: Where Have We Been and Where Are We Going?, 2 J. HEALTH & LIFE SCI. L. 73, 100-01 (2009)

[408] W. Nicholson Price II, Making Do in Making Drugs: Innovation Policy and Pharmaceutical Manufacturing, 55 B.C. L. REV. 491, 525 n.229 (2014)

[409] Daniel R. Goldberg. 'Aspirin: Turn of the Century Wonder Drug.' Distillations. Summer 2009. Accessed on: 28 September 2018. https://www.sciencehistory.org/distillations/magazine/aspirin-turn-of-the-century-miracle-drug

[410] Scherer, Matthew U. 'Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies.' Harv. JL & Tech. 29 (2015): 353.

[411] Aviation Accident Reports. National Transportation Safety Board. Accessed on: 28 September 2018 https://www.ntsb.gov/investigations/AccidentReports/Pages/aviation.aspx [https://perma.cc/US7N-3UCR]

[412] Jack Smith IV, 'Crime-prediction tool PredPol amplifies racially biased policing, study shows,' Mic, Oct. 9, 2016, Accessed on: 28 September 2018. https://mic.com/articles/156286/crime-prediction-tool-pred-pol-only-amplifies-racially-biased-policing-study-shows

[413] Andrew G. Ferguson, The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement, (New York: NYU Press, 2017).

[414] James Vincent, 'Google uses DeepMind AI to cut data center energy bills,' The Verge, July 21, 2016, https://www.theverge.com/2016/7/21/12246258/google-deepmind-ai-data-center-cooling

[415] Dutchnews, Dutch councils use algorithms to identify potential social security fraudsters, Dutchnews.nl, April 9, 2018, https://www.dutchnews.nl/news/2018/04/dutch-councils-use-algorithms-to-identify-potential-social-security-fraudsters/.

[416] Kade Crockford, 'Risk assessment tools in the criminal justice system: inaccurate, unfair, and unjust?,' ACLU of Massachusetts, March 8, 2018, https://privacysos.org/blog/risk-assessment-tools-criminal-justice-system-inaccurate-unfair-unjust

[417] Virginia Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, (New York: St. Martin's Press, 2018); Nazgol Ghandnoosh, Black Lives Matter: Eliminating Racial Inequity in the Criminal Justice System (Washington DC: The Sentencing Project, 2015), http://sentencingproject.org/wp-content/uploads/2015/11/Black-Lives-Matter.pdf

[418] Insha Rahman, 'The State of Bail: A Breakthrough Year for Bail Reform,' Vera Institute of Justice, 2017, https://www.vera.org/state-of-justice-reform/2017/bail-pretrial

[419] Dillon Reisman, Jason Schultz, Kate Crawford, Meredith Whittaker, 'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability', AI Now, April 2018 https://ainowinstitute.org/aiareport2018.pdf

[420] Ali Winston, 'Transparency Advocates Win Release of NYPD 'Predictive Policing' Documents,' The Intercept, Jan. 27, 2018, https://theintercept.com/2018/01/27/nypd-predictive-policing-documents-lawsuit-crime-forecasting-brennan/ .

[421] Leonard Ortolano and Anne Shepard, 'Environmental impact assessment: challenges and opportunities,' Impact assessment 13, no. 1 (1995): 3-30. https://www.tandfonline.com/doi/abs/10.1080/07349165.1995.9726076

[422] United Nations, 'Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework,' 20-24 (2011), http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

[423] Kenneth A. Bamberger and Deirdre Mulligan, 'Privacy decision-making in Administrative authorities,' Chicago L. Rev. 75(1):75 (2008), https://www.truststc.org/pubs/258.html

[424] 'Data Protection Impact Assessments,' Information Commissioner's Office, accessed March 16, 2018, https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/

[425] 'Data protection impact assessment,' Art. 35, Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1.

[426] Catherine Crump, 'Surveillance Policy Making by Procurement,' Wash. L. Rev. 91 (2016): 1595.

[427] European Commission, Commission Recommendation 2017/1805, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017H1805&from=EN

[428] Atila Abdulkadiroğlu, Yeon-Koo Che, and Yosuke Yasuda, 'Expanding' choice' in school choice,' American Economic Journal: Microeconomics 7, no. 1 (2015): 1-42;

[429] Neil Thakral, 'The Public-Housing Allocation Problem,' Technical report, Harvard University, 2016.

[430] 'Review of quality assurance of Government analytical models: final report,' HM Treasury, UK, March 2013, https://www.gov.uk/government/publications/review-of-quality-assurance-of-government-models).

[431] Aaron Reike, Miranda Bogen and David G. Robinson, Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods, (Upturn and Omidyar Network, 2018), https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods

[432] April Glaser, 'Who Trained Your A.I.,' Slate, Oct. 24, 2017, http://www.slate.com/articles/technology/technology/2017/10/what_happens_when_the_data_used_to_train_a_i_is_biased_and_old.html.

[433] Batya Friedman and Helen Nissenbaum, 'Bias in computer systems,' ACM Transactions on Information Systems (TOIS) 14, no. 3 (1996): 330-347. https://www.nyu.edu/projects/nissenbaum/papers/biasincomputers.pdf.

[434] Steven L. Chanenson and Jordan M. Hyatt, 'The Use of Risk Assessment at Sentencing: Implications for Research and Policy,' Villanova Law/Public Policy Research Paper No. 2017-1040 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2961288.

[435] 'Definitions,' Art. 4, Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1.

[436] U.S. Department of Energy, EERE Project Management Center, NEPA Determination. 2013, https://www.energy.gov/sites/prod/files/2013/06/f1/CX-010268.pdf

[437] U.S. Department of Agriculture, Forest Service, NEPA Categorical Exclusion Checklist. 2014, https://www.fs.usda.gov/nfs/11558/www/nepa/97961_FSPLT3_1655236.pdf.

[438] New York City Council, Hearing Testimony, Oct. 16, 2017, http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0.

[439] The Federal Trade Commission. 'Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues.' January 2016. https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report.

[440] David McCabe, 'Lawmakers Are Trying to Understand How Tech Giants' Algorithms Work,' Axios, Nov. 29, 2017, https://www.axios.com/lawmakers-are-trying-to-understand-how-tech-giants-algorithms-work-1513307255-b4109efc-9566-4e69-8922-f37d9e829f1f.html

[441] Eric Holder, 'Speech at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference' (Philadelphia, PA, Aug. 1, 2014), Department of Justice, https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

[442] John Fry, Anne Maxwell, Sarah Apere, Paddy McAweeney, Luke McSharry, and Ainhoa Gonza lez, 'Non-Technical Summaries-Due Care and Attention,' In 34th IAIA Annual Conference, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.567.8444&rep=rep1&type=pdf.

[443] Saunders, et al., 'Predictions put into practice.'

[444] Danielle Keats Citron, 'Big Data Should Be Regulated by 'Technological Due Process,'' N.Y. Times, July 29, 2016, https://www.nytimes.com/roomfordebate/2014/08/06/is-big-data-spreading-inequality/big-data-should-be-regulated-by-technological-due-process.

[445] Citron DK., 'Technological Due Process.'; Citron & Pasquale, The Scored Society.'; Crawford and Schultz. 'Big data and due process.'

[446] Diakopolous, et al., 'Principles for Accountable Algorithms and a Social Impact Statement for Algorithms,' FATML, accessed March 16, 2018, https://www.fatml.org/resources/principles-for-accountable-algorithms.

[447] Erin Griffith, 'The Other Tech Bubble,' Wired, Dec. 16, 2017, https://www.wired.com/story/the-other-tech-bubble/.

[448] Katherine Fink, 'Opening the government's black boxes: freedom of information and algorithmic accountability, Information,' Communication & Society (2017), https://www.tandfonline.com/doi/pdf/10.1080/1369118X.2017.1330418.

[449] Nicholas Diakopoulos, 'We need to know the algorithms the government uses to make important decisions about us,' The Conversation, May 23, 2016, https://theconversation.com/we-need-to-know-the-algorithms-thegovernment-uses-to-make-important-decisions-about-us-57869.

[450] Kamira Laachir, 'France's 'Ethnic' Minorities and the Question of Exclusion,' Mediterranean Politics 12, no. 1 (March 2007): 99-105

[451] Frédérick Douzet & Jérémy Robine, ''Les jeunes des banlieues': neighborhood effects on the immigrant youth experience in France', Journal of Cultural Geography 32, No. 1(2015):40-53

[452] Clare Foran, 'How France Built Inequality Into Its Cities,' CityLab, November 20, 2012, https://www.citylab.com/equity/2012/11/how-france-built-discrimation-its-cities/3881/.

[453] Philippe Sotto, 'Top French court: police illegally checked 3 minority men,' Associated Press, November 09, 2016, https://www.apnews.com/9a55dab8c4fa46878a9edad27901925d.

[454] Solon Barocas, Kate Crawford, Aaron Shapiro and Hanna Wallach, 'The Problem with Bias: From Allocative to Representational Harms in Machine Learning', SIGCIS conference paper, October 2017

[455] Kate Crawford, 'The Trouble with Bias', NIPS conference keynote, December 2017, https://www.youtube.com/watch?v=fMym_BKWQzk.

[456] Tom Simonite, 'When it comes to gorillas, Google photos remains blind,' Wired, January 11, 2018, https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind.

[457] Kenneth R. Harney, 'Zip code 'redlining': a sweeping view of risk,' Washington Post, February 2, 2008, http://www.washingtonpost.com/wp-dyn/content/article/2008/02/01/AR2008020101680.html

[458] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 'Datasheets for Datasets.' arXiv preprint arXiv:1803.09010 (2018), https://arxiv.org/abs/1803.09010.

[459] Jessica Saunders, Priscillia Hunt, and John S. Hollywood, 'Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot,' Journal of Experimental Criminology 12:3 (2016), 347-371, https://link.springer.com/article/10.1007/s11292-016-9272-0 .

[460] Danielle Keats Citron, 'Big Data Should Be Regulated by 'Technological Due Process,'' N.Y. Times, July 29, 2016, https://www.nytimes.com/roomfordebate/2014/08/06/is-big-data-spreading-inequality/big-data-should-be-regulated-by-technological-due-process

[461] Wexler, 'Life, Liberty, and Trade Secrets'; Ram, 'Innovating Criminal Justice'.

[462] David S. Levine, 'The People’s Trade Secrets,' 18 Mich. Telecomm. & Tech. L. Rev. 61 (2011), https://repository.law.umich.edu/mttlr/vol18/iss1/2/.

[463] Jan Whittington, Ryan Calo, Mike Simon, and Jesse Woo, 'Push, Pull, and Spill: A Transdisciplinary Case Study In Municipal Open Government,' 30 Berkeley Tech. L.J. 1967 (2015), https://scholarship.law.berkeley.edu/btlj/vol30/iss2/2/ .

[464] Mike Ananny and Kate Crawford, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability.' New Media & Society (2016).

[465] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort, 'Auditing algorithms: Research methods for detecting discrimination on internet platforms,' Data and discrimination: converting critical concerns into productive inquiry (2014): 1-23.

[466] Devin G. Pope and Justin R. Sydnor. 'Implementing anti-discrimination policies in statistical profiling models.' American Economic Journal: Economic Policy 3, no. 3 (2011): 206-31.

[467] Kristian Lum and William Isaac, 'To predict and serve?,' Significance 13, no. 5 (2016): 14-19. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2016.00960.x/full

[468] Conference on Fairness, Accountability, and Transparency, https://fatconference.org .

[469] AI Now 2016 Symposium, July 7, 2016, https://ainowinstitute.org/events/2016-symposium.html;

[470] AI Now 2017 Symposium, July 10, 2017, https://ainowinstitute.org/events/2017-symposium.html.

[471] Executive Office of the President, Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights, May 2016, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

[472] U.S. Department of Justice, Office of the Inspector General (2018), https://oig.justice.gov.

[473] Phoebe Friesen, Lisa Kearns, Barbara K. Redman and Arthur L. Caplan, 'Extending Ethical Strides: From Tribal IRBs to the Bronx Community Research Review Board,' The American Journal of Bioethics (2017), 17:11, W5-W8, https://www.tandfonline.com/doi/abs/10.1080/15265161.2017.1378755.

[474] The 'checking the box' mentality is a common critique of workplace sexual harassment training (Yuki Noguchi, 'Trainers, Lawyers Say Sexual Harassment Training Fails,' All Things Considered, NPR, Nov. 8, 2017, https://www.npr.org/2017/11/08/562641787/trainers-lawyers-say-sexual-harassment-training-fails).

[475] Rebecca Wexler, 'Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System,' 70 Stan. L. Rev., (forthcoming 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2920883

[476] Natalie Ram, 'Innovating Criminal Justice,' Northwestern L. Rev. (forthcoming 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3012162 .

[477] Nicole A. Ozer, 'Santa Clara County Passes Landmark Law to Shut Down Secret Surveillance,' ACLU of Northern California, June 8, 2016, https://www.aclunc.org/blog/santa-clara-county-passes-landmark-law-shut-down-secret-surveillance).

[478] 'National Environmental Policy Act Review Process,' US Environmental Protection Agency, accessed June 21, 2018, https://www.epa.gov/nepa/national-environmental-policy-act-review-process.

[479] Directive (EU) 2014/24, of the European Parliament and of the Council of 26 February 2014 on public procurement (The Public Contracts Directive 2014);

[480] Directive (EU) 2014/23, of the European Parliament and of the Council of 26 February 2014 on the award of concession contracts (The Concessions Contracts Directive 2014);

[481] Directive (EU) 2014/25, of the European Parliament and of the Council of 26 February 2014 on procurement by entities operating in the water, energy, transport and postal services sectors (The Utilities Directive 2014).

[482] ISO 9000 family - Quality management. International Standards Organization (ISO). Accessed on: 4 October 2018. https://www.iso.org/iso-9001-quality-management.html

[483] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. IEEE Standards Association. Accessed on: 4 October 2018. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[484] ISO/IEC JTC 1/SC 42 Artificial Intelligence. International Standards Organization (ISO). Accessed on: 28 September 2018. https://www.iso.org/committee/6794475.html

[485] Guynn, Jessica. 'Google Photos labeled black people 'gorillas''. *USA Today*, 01 July 2015. Accessed on: 04 October 2018. https://eu.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/

[486] Hill, Kashmir. 'How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did'. *Forbes,* 12 February 2012. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

[487] Algorithms – censorship à la carte?. EDRi, 12 July 2016. https://edri.org/algorithms-censorship-a-la-carte/

[488] Freeman, Katherine. 'Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis.' *North Carolina Journal of Law and Technology* 18 (2016): 75-180.

[489] RightsCon Toronto 2018. Toronto, Canada. https://rightscon2018.sched.com/

[490] The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems. *AccessNow,* 16 May 2018. RightsCon. Toronto, Canada. https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/

[491] Evidence submitted by The Human Rights, Big Data and Technology Project, available at <http://www.hrbdt.ac.uk> to House of Lords inquiry on Ariticial Intelligence http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69717.html

[492] https://www.parliament.uk/ai-committee

[493] V Ng and C Kent, Human Rights in the Digital Age: The Perils of Big Data and Technology (Part II) (2016) available at: www.hrbdt.ac.uk/human-rights-in-the-digital-age-the-perils-of-big-data-and-technology-part-ii/ [last accessed 06.09.17].

[494] OHCHR, A Human Rights Based Approach to Data: Leaving No One Behind in the 2030 Development Agenda: Guidance Note to Data Collection and Disaggregation (2016) available at: <http://hrbaportal.org/wp-content/files/GuidanceNoteonApproachtoData.pdf> [last accessed 05.09.17].

[495] T Harris and J Wyndham, 'Data Rights and Responsibilities: A Human Rights Perspective on Data Sharing' (2015) 10(3) Journal of Empirical Research on Human Research Ethics 334 – 337

[496] Office of the United Nations High Commissioner for Human Rights. Frequently Asked Questions on a Human Rights-Based Approach to Development Cooperation (United Nations, New York and Geneva, 2006) available at: www.ohchr.org/Documents/Publications/FAQen.pdf [last accessed on 05.09.17] pg. 17.

[497] The Human Rights-Based Approach to Development Cooperation: Towards a Common Understanding Among the United Nations Agencies (Second Inter-Agency Workshop, Stamford, USA, May 2003) available in Annex II at: www.ohchr.org/Documents/Publications/FAQen.pdf [last accessed on 05.09.17]

[498] Harrison, James. 'Human rights measurement: reflections on the current practice and future potential of human rights impact assessment.' Journal of Human Rights Practice 3, no. 2 (2011): 162-187.

[499] Schwab, Klaus. 'The 4th industrial revolution.' In World Economic Forum. New York: Crown Business. 2016.

[500] Dutton, Tim. 'An Overview of National AI Strategies'. *Medium,* 28 June 2018. https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

[501] Hughes, Mark. 'Artificial intelligence is now an arms race. What if the bad guys win?'. *World Economic Forum*, 10 November 2017. https://www.weforum.org/agenda/2017/11/cybersecurity-artificial-intelligence-arms-race/

[502] https://www.ft.com/content/856753d6-8d31-11e7-a352-e46f43c5825d

[503] Tomasik, Brian. 'International Cooperation vs. AI Arms Race'. *Foundational Research Institute*, 5 December 2013, updated 29 February 2016. https://foundational-research.org/international-cooperation-vs-ai-arms-race/

[504] Horowitz, Michael C. 'Artificial Intelligence, International Competition, and the Balance of Power (May 2018).' Texas National Security Review (2018).

[505] Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe et al. 'The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.' arXiv preprint arXiv:1802.07228 (2018).

[506] Stevens, Tim. 'Cyberweapons: an emerging global governance architecture.' (2017). Palgrave Communications volume 3, Article number: 16102 (2017) https://doi.org/10.1057/palcomms.2016.102

[507] Sparrow, Robert. 'Predators or plowshares? Arms control of robotic weapons.' IEEE Technology and Society Magazine 28, no. 1 (2009): 25-29.

[508] Cave, Stephen, and Seán SOhÉigeartaigh. 'An AI Race for Strategic Advantage: Rhetoric and Risks.' In AAAI/ACM Conference on Artificial Intelligence, Ethics and Society. 2018.

[509] Chung, Lawrence, Brian A. Nixon, Eric Yu, and John Mylopoulos. Non-functional requirements in software engineering. Vol. 5. Springer Science & Business Media, 2012.

[510] Downes, Larry. 'GDPR and the End of the Internet's Grand Bargain.' (2018).

[511] Mishra, Neha. 'Data Localization Laws in a Digital World: Data Protection or Data Protectionism?.' (2015).

[512] https://iapp.org/news/a/brazil-moving-forward-with-gdpr-inspired-data-protection-bill/

[513] https://www.computerworlduk.com/data/how-chinas-data-privacy-law-was-inspired-by-gdpr-3678918/

[514] https://www.jdsupra.com/legalnews/new-gdpr-inspired-data-laws-in-brazil-95445/

[515] Botha, Johnny, M. M. Grobler, Jade Hahn, and Mariki Eloff. 'A High-Level Comparison Between the South African Protection of Personal Information Act and International Data Protection Laws.' In ICMLG2017 5th International Conference on Management Leadership and Governance, p. 57. Academic Conferences and publishing limited, 2017.

[516] https://iapp.org/news/a/brazil-moving-forward-with-gdpr-inspired-data-protection-bill/

[517] https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/108/signatures

[518] https://www.coe.int/en/web/data-protection/convention108/modernised

[519] Greenleaf, Graham. 'The influence of European data privacy standards outside Europe: implications for globalization of Convention 108.' International Data Privacy Law 2, no. 2 (2012): 68-92.

[520] Stackowiak, Robert. 'Why Diversity Matters.' In Remaining Relevant in Your Tech Career, pp. 41-52. Apress, Berkeley, CA, 2019.

[521] Zou, James, and Londa Schiebinger. 'AI can be sexist and racist—it's time to make it fair.' Nature 559, 324-326 (2018) [doi: 10.1038/d41586-018-05707-8]

[522] https://www.forbes.com/sites/kalevleetaru/2016/05/13/is-facebooks-trending-topics-biased-against-africa-and-the-middle-east/#3c345e1463db

[523] Woolley, Samuel C. 'Automating power: Social bot interference in global politics.' First Monday 21, no. 4 (2016).

[524] Forelle, Michelle, Phil Howard, Andrés Monroy-Hernández, and Saiph Savage. 'Political bots and the manipulation of public opinion in Venezuela.' arXiv preprint arXiv:1507.07109 (2015).

[525] Matz, S. C., Michal Kosinski, Gideon Nave, and David J. Stillwell. 'Psychological targeting as an effective approach to digital mass persuasion.' Proceedings of the national academy of sciences 114, no. 48 (2017): 12714-12719.

[526] Forelle, Michelle, Phil Howard, Andrés Monroy-Hernández, and Saiph Savage. 'Political bots and the manipulation of public opinion in Venezuela.' arXiv preprint arXiv:1507.07109 (2015).

[527] Howard, Philip N., Samuel Woolley, and Ryan Calo. 'Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration.' Journal of Information Technology & Politics 15, no. 2 (2018): 81-93.

[528] Lin, Herbert S. 'Offensive cyber operations and the use of force.' J. Nat'l Sec. L. & Pol'y 4 (2010): 63.

[529] Patton, Clifford A. 'The Impact of Cyber Espionage: Changing Perceptions with the US vis-a-vis the Transatlantic.' PhD diss., 2017.

[530] Blank, Stephen. 'Cyber War and Information War à la Russe.' Understanding Cyber Conflict. 14 Analogies (2017): 81-98.

[531] Hamilton, Logan. 'Beyond Ballot-Stuffing: Current Gaps in International Law regarding Foreign State Hacking to Influence a Foreign Election.' Wis. Int'l LJ 35 (2017): 179.

[532] Lam, Christina. 'A Slap on the Wrist: Combatting Russia's Cyber Attack on the 2016 US Presidential Election.' BCL Rev. 59 (2018): 2167.

[533] Buse, Mihaiela. 'EUROPEAN UNION CYBER SECURITY IN A GLOBALIZED WORLD.' In International Scientific Conference' Strategies XXI', vol. 1, p. 159. ' Carol I' National Defence University, 2017.

[534] Väisänen, Teemu, Christian Braccini, Michael Sadloň, Hayretdin Bahşi, Agostino Panico, and Kris van der Meij. 'Battlefield Digital Forensics: Digital Intelligence and Evidence Collection in Special Operations.' (2016).

[535] Common, MacKenzie F. 'Facebook and Cambridge Analytica: let this be the high-water mark for impunity.' LSE Business Review (2018).

[536] https://www.tralac.org/images/docs/12964/wto-general-council-joint-statement-on-electronic-commerce-initiative-proposal-for-the-exploratory-work-by-japan-april-2018.pdf

[537] https://www.tralac.org/images/docs/12964/wto-general-council-joint-statement-on-electronic-commerce-initiative-communication-from-the-united-states-april-2018.pdf

[538] Fung, Archon, Mary Graham, and David Weil. Full Disclosure: The Perils and Promise of Transparency. Cambridge University Press, 2009

[539] https://www.itu.int/en/ITU-T/AI/Pages/default.aspx

[540] Copeland, Damian, and Luke Reynoldson. 'How to avoid'summoning the demon': The legal review of weapons with artificial intelligence.' Pandora's Box 2017 (2017): 97.

[541] https://en.unesco.org/news/unesco-s-communication-and-information-sector-invites-experts-contribute-open-discussion

[542] https://rm.coe.int/algorithms-and-human-rights-study-on-the-human-rights-dimension-of-aut/1680796d10

[543] https://g7.gc.ca/en/g7-presidency/themes/preparing-jobs-future/g7-ministerial-meeting/chairs-summary/annex-b/

[544] Twomey, Paul. Building on the Hamburg Statement and the G20 Roadmap for Digitalization: Toward a G20 framework for artificial intelligence in the workplace. No. 2018-63. Economics Discussion Papers, 2018.

[545] http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/

[546] Vienna Declaration and Programme of Action [Adopted by the World Conference on Human Rights in Vienna on 25 June 1993] available at: <www.ohchr.org/EN/ProfessionalInterest/Pages/Vienna.aspx> [last accessed on 06.09.17]

[547] Office of the United Nations High Commissioner for Human Rights, The Core International Human Rights Instruments and their Monitoring Bodies (undated) available at: <www.ohchr.org/EN/ProfessionalInterest/Pages/CoreInstruments.aspx> [last accessed 06.09.17].

[548] OECD Artificial Intelligence Expert Group (AIGO) http://www.oecd.org/going-digital/ai/

[549] Global Governance of AI Roundtable at the World Government Summit in Dubai https://www.worldgovernmentsummit.org/

[550] Council of Europe and Artificial Intelligence https://www.coe.int/AI

[551] Erdelyi, Olivia Johanna and Goldsmith, Judy, Regulating Artificial Intelligence: Proposal for a Global Solution (February 2, 2018). 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2--3, 2018, New Orleans, LA, USA doi/10.1145/3278721.3278731. Available at SSRN: https://ssrn.com/abstract=3263992

[552] Wallach, Wendell and Marchant, Gary E., An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics (February 2, 2018). 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2--3, 2018, New Orleans, LA, USA

[553] Wilson, Dave, and Joseph S. Valacich. 'Unpacking the privacy paradox: Irrational decision-making within the privacy calculus.' (2012).

[554] Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 'Fair, Transparent, and Accountable Algorithmic Decision-making Processes.' Philosophy & Technology (2017): 1-17.

[555] Klawitter, Erin, and Eszter Hargittai. ''It's Like Learning a Whole Other Language:' The Role of Algorithmic Skills in the Curation of Creative Goods.' International Journal of Communication 12 (2018): 21.

[556] boyd, d., Crawford, K.: Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication, & Society 15(5), 662{679 (2012)

[557] Bhargava, R., Deahl, E., Letouze, E., Noonan, A., Sangokoya, D., Shoup, N.: Beyond data literacy: Reinventing community engagement and empowerment in the age of data. Data-Pop Alliance White Paper Series (2015). URL http://datapopalliance.org/wp-content/uploads/2015/11/Beyond-Data- Literacy-2015.pdf

[558] Burrell, J. (2016). How the machine `thinks': Understanding opacity in machine learning algorithms. Big Data & Society 3(1)

[559] Elvira Perez Vallejos, Ansgar Koene, Virginia Portillo, Liz Dowthwaite and Monica Cano, 'Young people's policy recommendations on algorithm fairness', in Proceedings of the 2017 ACM on Web Science Conference, Pages 247-251 ISBN: 978-1-4503-4896-6 DOI: 10.1145/3091478.3091512

[560] Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., . . . Sandvig, C. (2015). '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 (pp. 153–162). New York, NY: Association for Computing Machinery.

[561] Papsdorf, Christian, and Sebastian Jakob. 'Ein Kampf gegen Windmühlen: Jugendliche und junge Erwachsene im Umgang mit Algorithmen und Überwachung im Internet.' kommunikation@ gesellschaft 18 (2017): 27.

[562] Moses, Lyria Bennett. 'Is Your Algorithm Dangerous?[Leading Edge].' IEEE Technology and Society Magazine 37, no. 3 (2018): 20-21.

[563] Baker, Jamie J. 'Beyond the Information Age: The Duty of Technology Competence in the Algorithmic Society.' SCL Rev. 69 (2017): 557.

[564] Buchanan, B., and Taylor, M. 2017. 'Machine Learning for Policymakers,' Paper, Cyber Security Project, Belfer Center. https://www.belfercenter.org/publication/machine-learning-policymakers

[565] Naughton, John. 'Death by drone strike, dished out by algorithm'. *The Guardian,* 21 February 2016. https://www.theguardian.com/commentisfree/2016/feb/21/death-from-above-nia-csa-skynet-algorithm-drones-pakistan

[566] Nick Hopkins, 'Revealed: Facebook's internal rulebook on sex, terrorism and violence,' The Guardian, May 21, 2017

[567] Mike Isaac, 'How Uber Deceives the Authorities Worldwide,' The New York Times, Mar. 3, 2017

[568] Mike Isaac, 'Justice Department Expands Its Inquiry Into Uber's Greyball Tool,' The New York Times, Mar. 5, 2017. 39

[569] Chouldechova, Alexandra. 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.' Big data 5, no. 2 (2017): 153-163.

[570] Wadsworth, Christina, Francesca Vera, and Chris Piech. 'Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction.' arXiv preprint arXiv:1807.00199 (2018).

[571] Flores, Anthony W., Kristin Bechtel, and Christopher T. Lowenkamp. 'False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.' Fed. Probation 80 (2016): 38.

[572] Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 'Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment.' In Proceedings of the 26th International Conference on World Wide Web, pp. 1171-1180. International World Wide Web Conferences Steering Committee, 2017.

[573] Howard, Alexander Benjamin. 'The art and science of data-driven journalism.' (2014).

[574] Eilam, Eldad. Reversing: Secrets of Reverse Engineering. Wiley, 2005

[575] https://ec.europa.eu/commission/news/whistleblower-protection-2018-apr-23_en

[576] Mateski, Mark, Cassandra M. Trevino, Cynthia K. Veitch, John Michalski, J. Mark Harris, Scott Maruoka, and Jason Frye. 'Cyber threat metrics.' Sandia National Laboratories (2012).

QA-03-19-162-EN-N-

Beginning with an analysis of the social, technical and regulatory challenges posed by algorithmic systems, this study explores policy options for the governance of algorithmic transparency and accountability. An extensive review and analysis of existing proposals for algorithmic system governance points to four policy options, addressing awareness raising, accountability in public sector use, regulatory oversight, and global coordination for algorithmic governance.

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

ISBN 978-92-846-4656-2 | doi:10.2861/59990 | QA-03-19-162-EN-N