algo:aware

Raising awareness on algorithms

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

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

Executive Summary

December 2018

Context

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

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

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

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

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

Introduction

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

The widespread application of algorithmic decision-making systems has been enabled by advancements in computing power and the increased ability to collect, store and utilise massive quantities and a variety of personal and non-personal data from both traditional and non-traditional sources. Algorithmic systems are capable of integrating more sources of data, and identifying relationships between those data, more effectively than humans can. In particular, they may be able to detect rare outlier cases where humans cannot.

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

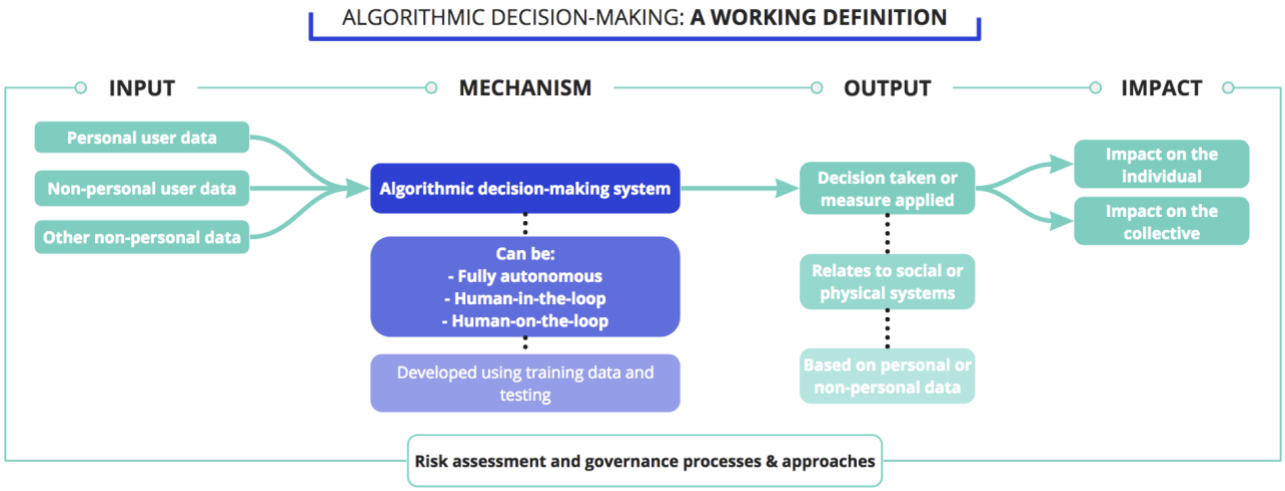
Thus, **real tensions exist between the positive impacts and the risks presented by of algorithmic decision-making** in both current and future applications. In the European Union, a regulatory framework already governs some of these concerns. The General Data Protection Regulation establishes a set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

Scope of the report

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

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

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



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

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

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

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

Mapping the Academic Debate

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

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

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

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

- **Liability** – questions of liability frequently arise in discussions about computational systems which have direct physical effects on the world (for instance self-driving cars).

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

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

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

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

- What definitions of fairness are appropriate and necessary for different instances of algorithmic decisions? What are the tradeoffs between them? What are the fairness benchmarks for specific algorithmic decisions and in what situations should algorithms be held to a greater standard of fairness than human decisions? What governance can establish and enforce such standards? Do citizens and businesses feel that systems which have been 'debiased' are more legitimate on the ground, and do such systems actually mitigate or reduce inequalities in practice?

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

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

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

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

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

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

Liability. The report presents the different legal models of liability and responsibility around algorithmic systems, including strict liability, negligence-based liability, and alternative reparatory policy approaches based on insurance schemes. It further explains situations where court cases have attributed liability for defamatory content on search engines.

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

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

There is significant effort being directed towards tackling the challenges facing algorithmic decision-making by industry, civil society, academia and other interested parties. This is true across all categories of initiatives examined and relates to all of the perspectives discussed

above. In particular, there are a large number of initiatives aimed at promoting responsible decision-making algorithms through codes of conduct, ethical principles or ethical frameworks.

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

- **Standardisation efforts:** ISO and the IEEE are two of the most prominent global standards bodies, with the buy-in and cooperation of a significant number of national standards bodies. As such, it is important that these organisations are working towards tackling a number of these challenges. The final effort documented here, outside of the scope of the ISO and the IEEE, is the Chinese White Paper on Standardisation. Although no concrete work has been conducted, this document illustrates that stakeholders currently involved in the standardisation process in China – a multi-disciplinary group – are considering algorithmic decision-making from all the key perspectives being discussed.
- **Codes of conduct, ethical principles and frameworks:** As mentioned above, there are a vast number of attempts to govern the ethics of AI development and use with no clear understanding or reporting on take-up or impact. These initiatives have been initiated by stakeholders from all relevant groups, in some cases in isolation but also through multi-disciplinary efforts. Furthermore, much of this work attempts to tackle the challenges facing algorithmic decision-making from multiple perspectives. For instance, the ethical principles developed by the Software and Information Industry Association (SIIA) explicitly discuss the need for transparency and accountability; and the Asilomar Principles, which cover, in particular, topics of fairness, transparency, accountability, robustness and privacy. Interesting work that stands out and could be beneficial on a higher plane includes the work of Algorithmenethik on determining the success factors for a professional ethics code and the work of academics Cowls and Floridi, who recognised the emergence of numerous codes with similar principles and conducted an analysis across some of the most prominent examples. Cowls and Floridi's work is also valuable as it ties the industry of AI development and algorithmic decision-making to long established ethical principles from bioethics. The elements of learning these examples bring from established sectors can be extremely useful.
- **Working groups and committees:** The initiatives examined have primarily been initiated by civil society organisations (including, for example, AlgorithmWatch and the Machine Intelligence Research Institute) with the aim of bringing together a wide variety of stakeholders. Outputs of these initiatives tend to include collaborative events, such as the FAT/ML workshops, or research papers and advice, such as the World Wide Web Foundation's white paper series on *Opportunities and risks in emerging technologies*. As for the above, this type of initiative is often focused on tackling the challenges facing algorithmic decision-making from multiple perspectives. For instance, AlgorithmWatch maintains scientific working groups, which, in the context of various challenges, discuss, amongst others, topics of non-discrimination and bias, privacy and algorithmic robustness. Furthermore, no clear information on the impact of these initiatives is currently available.
- **Policy and technical tools:** In this category, the initiatives examined have been developed by academic research groups (e.g. the work of NYU's AI Now Institute and the UnBias research project), civil society (e.g. the Digital Decisions Tool of the Center for Democracy and Technology) or multi-disciplinary groups (e.g. the EthicsToolkit.ai

developed through collaboration between academia and policymakers). In terms of how these tools address the challenges facing algorithmic decision-making, they tend to focus on specific challenges; a clear example being the 'Fairness Toolkit', developed by the UnBias research project.

Policy initiatives and approaches

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

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

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

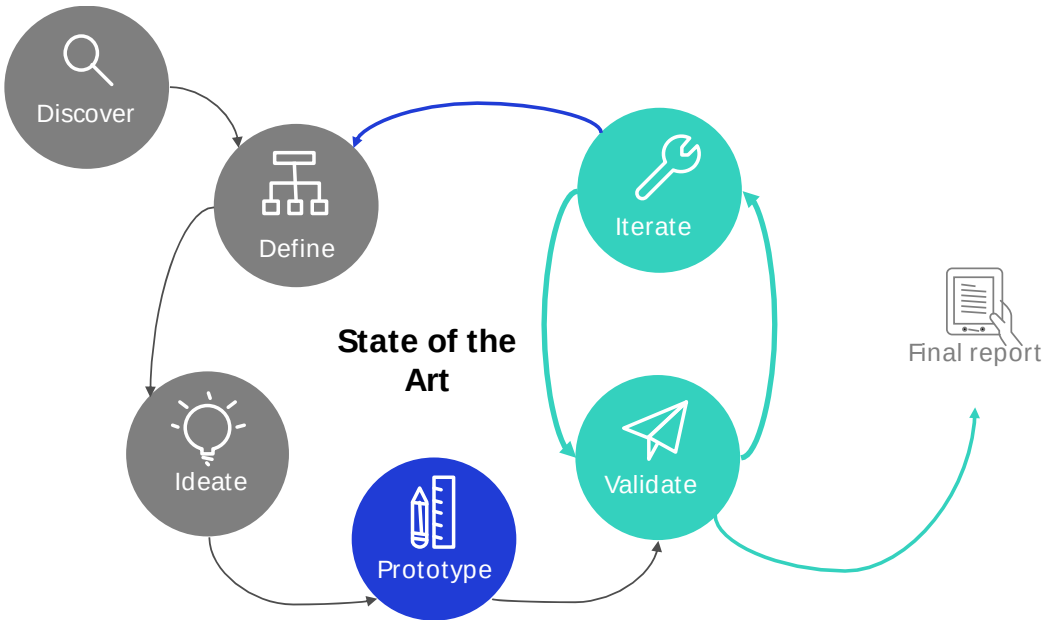
More commonly, however, the initiatives are softer in nature. These initiatives also reflect the aim of harnessing the potential of AI through the development of wide-reaching industrial and research strategies. Prominent types of initiatives implemented globally include:

- Development of **strategies on the use of AI and algorithmic decision-making**, with examples including France's *AI for Humanity* Strategy, which focuses on driving AI research, training and industry in France alongside the development of an ethical framework for AI to ensure, in particular, transparency, explainability and fairness. Another example is the Indian *National AI Strategy* and the EUR 3bn AI strategy issued by Germany in November 2018, which aims at making the country a frontrunner in the second AI wave, while maintaining strong ethical principals. Related to this are the numerous White Papers and reports developed, including the *German White Paper on AI*, the *Visegrád position paper on AI* and the Finnish *Age of AI* report.

- Establishment of **expert groups and guidance bodies** with examples including the Group of Experts and "Sages" established in Spain in 2018, the Italian *AI Task Force* and the German *Enquete Commission*. Considering the former example, this group has been tasked with guiding on the ethics of AI and Big Data through an examination of the social, juridicial and ethical implications of AI.

Next steps

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



algo:aware is procured by the European Commission and delivered by Optimity Advisors.

algo:aware *aims to assess the opportunities and challenges that emerge where algorithmic decisions have a significant bearing on citizens and where they produce societal or economic effects which need public attention.*

www.optimityadvisors.com

www.twitter.com/optimityeurope
www.linkedin.com/company/optimityeurope

Study contact: Quentin Liger – quentin.liger@optimtyadvisors.com