Universität Zürich UZH
Universität Basel Juristische Fakultät
e-PIAF Electronic Public Institutions and Administrations Research Forum
CENTER FOR INFORMATION TECHNOLOGY SOCIETY AND LAW — ITSL

Comprehensible Algorithms: A Legal Framework for the Use of Artificial Intelligence

Workshop Report

by Florent Thouvenin, Stephanie Volz, Fabienne Graf and Soraya Weiner (all University of Zurich), Nadja Braun Binder and Liliane Obrecht (both University of Basel), as well as invited experts Annika Baumann (Weizenbaum-Institut), Yochanan Bigman (Hebrew University), Andrea Bonezzi (New York University), Mireille Hildebrandt (Radboud University), Sandra Matz (Columbia University), Elena Mugellini (University of Applied Sciences of Western Switzerland, Fribourg), Massimiliano Ostinelli (Florida Atlantic University) and Matthias Spielkamp (Private Researcher, AlgorithmWatch).

From 30 March to 1 April 2023, a group of experts in the fields of law, business administration, business informatics, computer science, philosophy, and psychology discussed and investigated regulatory approaches to “Artificial Intelligence” (AI) at a workshop in Basel, Switzerland. The workshop was jointly organized by the Electronic Public Institutions and Administration Research Forum (e-PIAF) of the University of Basel and the Center for Information Technology, Society, and Law (ITSL) of the University of Zurich. The workshop is part of the joint research project “Comprehensible Algorithms: A Legal Framework for the Use of Artificial Intelligence”. The research project is funded by the Stiftung Mercator Schweiz. The workshop was funded by the Swiss National Science Foundation (SNSF) and the Freiwillige Akademische Gesellschaft Basel (FAG).

The participants engaged in lively and controversial discussions on various aspects of discrimination and manipulation caused by the use of “Artificial Intelligence”. Given that these topics are very broad and complex, and that time for discussion was limited, only some of the relevant issues were discussed. The resulting workshop summary highlights the main insights gained and shared by most participants. It is not meant to offer a comprehensive assessment of the issues related to discrimination and manipulation caused by the use of “Artificial Intelligence”. The main insights are the following:

1. Terminology: algorithmic systems

The term “Artificial Intelligence” can evoke misleading associations and diffuse fears. From a technical perspective, AI is an established collective term that encompasses a range of technologies that automate decisions, recommendations, conclusions, or predictions. AI includes knowledge-based systems, statistical methods, and machine learning approaches (e.g., neural networks). The high-performance use of these technologies is mainly based on the combination of a large number of mathematical optimizations, extracting structures from significant amounts of data, using large computing capacities.

To avoid misleading associations, the participants agreed not to use the term AI in this workshop summary but rather speak of “algorithmic systems.” This term does not refer to specific current or future technologies but to the application of technologies in a social context. The need for legal assessment arises when technologies affect individuals and/or society. The term “algorithmic systems” also covers applications with similar effects as what is now referred to as AI that are based on technologies not usually qualified as AI, as well as technologies yet to be developed.

2. Discrimination

2.1 Legal concept

Discrimination is a qualified mode of unequal treatment. Unequal treatment constitutes discrimination if less favorable treatment is grounded on membership of a certain group or a specific characteristic such as gender (incl. gender identity), social or ethnic origin, language, religion, political conviction, age, disability, and sexual orientation. Thus, discrimination is based on classifications according to a set of prohibited grounds. To qualify as discrimination, intent is not required.

Discrimination may be direct or indirect: Direct discrimination occurs when unequal treatment of a person is based on a prohibited ground. Indirect discrimination can be affirmed if a prima facie neutral characteristic has an impact that *de facto* disadvantages persons who share a particular attribute in relation to a prohibited ground. In computer science, a similar distinction is made between discrimination and proxy discrimination.

In principle, discrimination is prohibited. However, even if qualified grounds exist, discrimination may be justified. In these cases, Swiss doctrine and practice require an in-depth examination of the causes for the discriminatory treatment; based on such examination, discrimination can be justified if the discriminatory measure is a proportionate means to pursue a legitimate goal.
In Switzerland, the Federal Constitution obliges all state authorities to respect fundamental rights, including the prohibition of discrimination, and to contribute to their implementation (Art. 35 Federal

Constitution). In particular, they must ensure that the prohibition of discrimination, where applicable, also takes effect in relationships among private persons (Art. 8 para. 2 in conjunction with Art. 35 para. 3 Federal Constitution). Administrations and courts fulfill this duty by interpreting private law in light of the fundamental rights at stake. In contrast, there is no general prohibition of discrimination by private actors in Swiss law, only a rather specific provision in criminal law and two specific laws on gender equality and equality of persons with disabilities.

Against this background, the main research question discussed at the workshop was how discrimination through the use of algorithmic systems could be prevented. This question was addressed from different perspectives:

2.2 Preventing discrimination

a) Future forms of discrimination

While there are various ways to prevent discriminatory outcomes by algorithmic systems, a fundamental problem when trying to avoid discrimination is that it is not clear what will be considered discriminatory in the future. Consequently, an algorithmic system cannot be built and used to ensure it will indefinitely comply with legal requirements for non-discrimination.

b) Trade-off

Participants pointed out that in the design of algorithmic systems there can be a trade-off between maximizing predictive accuracy and minimizing bias that can lead to discrimination: While an algorithm can be designed in a way that minimizes bias, this may result in a reduction of the accuracy of the result.

c) Proxy discrimination

With machine learning methods, discrimination can result not only when prohibited grounds (e.g., race or gender) are directly used by the algorithm for the determination of an outcome (direct discrimination), but also when so-called proxies (e.g. a person's address instead of their race if a neighbourhood is predominantly inhabited by people of a certain race) correlate with prohibited grounds (indirect discrimination). Thus, in order to prevent algorithmic discrimination, it is not sufficient to ensure that no prohibited grounds are used by an algorithmic system. One participant believes that the complexities of deep learning will make it difficult if not impossible to decide which feature is a proxy of what other feature. As this defines the difference between direct and indirect discrimination, the consequences of such inability may be huge.

2.3 Detecting discrimination

a) Prohibited grounds

Several participants mentioned that it may be important that prohibited grounds are collected and used by the algorithmic system, as they are needed to detect discrimination and – ideally – counteract the problem.

b) Two-step test

The participants agreed that, in principle, two steps are necessary to detect discrimination. Firstly, unequal treatment has to be detected, and secondly, it has to be established whether the unequal treatment qualifies as discrimination. For implementing the first step, various approaches were discussed:

- The company that uses algorithmic systems could be obliged (by law) to document the design and deployment of the system, including the training data. If a complaint alleging discrimination is filed against the company, the company would have to provide documentation for an external audit.
- A “human-in-the-loop” could be instated to monitor the system while it is processing. This monitoring would usually be based on the concept of “explainable AI” and executed by an expert who understands the technical aspects of the system and the criteria that lead to discrimination.
- Also, an external expert could be appointed who would, in cases of a complaint, assess the problem. If the use of the system by the company alleged of discrimination, so far has produced too few cases to verify whether the algorithmic system discriminates, the external expert could be granted the right to request access to identical or sufficiently similar algorithmic systems and the corresponding data used by other companies.

- Other participants noted that some cases of discrimination may be difficult to detect, because the correlations are part of complex mathematical functions. This requires standard testing instead of waiting for a complaint.

3. Manipulation

A problem often associated with the use of algorithmic systems is the fear that individuals are manipulated in their thinking and behavior by such systems. As opposed to discrimination, which is a relatively well-researched and agreed-upon problem, the phenomenon of manipulation remains unclear, both from an empirical and from a normative perspective. This makes it especially hard to identify suitable regulatory approaches to address the problem of manipulation.

3.1 Examples

In order to better understand the phenomenon of manipulation, we first identified examples that could be considered problematic and, thus, legally relevant instances of manipulation.

- *Search engines:* systems that search for and identify items in a database that correspond to input data by the user. Results are often listed in a non-transparent order.

- *Social media:* applications and websites which are used for sharing and consuming various content (e.g., news, videos) as well as social interactions.
- *News feeds:* services that continuously deliver and update the latest news. They can be accessed, e.g., in newspaper applications or on social media platforms.
- *Targeted advertising:* advertising based on traits of the targeted consumer, mostly in online marketing. Consumer traits can contain information about personal preferences, demographics, and behavior.
- *Product recommendations:* photos of products and links to the platform where the products can be obtained are often displayed visually close to the primary content on websites and in applications. These recommendations can be personalized according to the consumers' traits.
- *Personalized pricing:* pricing models which charge individual consumers different prices for the same service or product (e.g., people who browse with more expensive devices have to pay more).
- *Dating apps:* applications that enable people to connect online to meet with potential romantic partners. Users are getting matched, e.g., according to their interests, age groups and geolocation. From a user perspective, bots are hard to detect.
- *Autofill:* text is getting automatically completed if users start typing. The generated text might imply new connections and information to the initial input data.
- *Infinite scrolling:* on websites or in applications, new content is displayed continuously when users scroll through it.

3.2 Regulatory approaches

Most participants agreed that the problem of manipulation cannot be encompassed by a singular regulatory approach and ought to be addressed with a suitable mix of legal instruments.

a) Prohibition of certain practices

One way of capturing instances of manipulation is by prohibiting certain uses of algorithmic systems *ex-ante*. While some participants favored a rather broad approach, prohibiting certain business models (e.g., using behavioral data for advertising), others were more skeptical and favored an *ex post* approach, combined with rather specific prohibitions (e.g., banning endless scrolling in social media apps used by children). Participants agreed that prohibitions might be needed to protect vulnerable groups such as children and persons with disabilities.

b) General clause

Given that the technology, the application of algorithmic systems, and the instances of manipulation are hard to define and will change rapidly over time, capturing all instances of manipulation by a single provision or a set of specific provisions seems very hard. Instead, general clauses in the existing law should be used to capture problematic instances of manipulation. In Switzerland, such clauses are namely found for the protection of personality (Art. 28 Civil Code) and with the general clause against unfair competition (Art. 2 Unfair Competition Act). The existing law lacks an agreed-upon understanding of the normative factors that allow one to distinguish between instances of legally problematic manipulation and cases of acceptable influencing. Instead of – or in addition to – identifying a set of normative criteria, one could also identify specific use cases that would be qualified as legally relevant manipulation.

c) Ensuring choice

Most participants agreed that algorithmic systems should be designed in a way to ensure that users are not automatically endorsing with the default decision proposed by the system. Ensuring choice may also mean that certain forms of choice architecture (i.e. "Dark Patterns") are considered unacceptable and prohibited. In addition, users should have a choice between personalized results and results based on non-personalized criteria, e.g., a chronological order of news.

d) Transparency of business models

A potential way to address manipulation is to make sure that experts (such as researchers, journalists, civil society organizations and supervisory bodies) have access to models, data and other relevant information that allows them to analyze and understand the technology used for a specific product or service as well as the underlying business model. This would allow them to inform the public and the individuals interacting with algorithmic systems to have a basic understanding of the technology and the business models.

e) Education and digital literacy

Manipulation of people is more difficult if people understand how an algorithmic system works. Therefore, educating people is an important way to fight manipulation. However, it must be ensured that this approach does not have the undesirable effect of shifting responsibility from companies to users ("You should have known that").

3.3 Normative criteria

Participants discussed various possible normative criteria that could be used to define legally problematic cases of manipulation.

a) Change of behavior

Participants agreed that a key feature of all instances of manipulation is a change of behavior of a single individual or a group of individuals that is caused by the interaction with an algorithmic system.

b) Harm

Most participants agreed that manipulation is only legally relevant if it causes some sort of harm. They agreed that the notion of harm should be broadly construed to cover (e.g.) physical harm, harm to mental health and financial damage, including loss of profit. However, others pointed out that this is only correct in a private law context, while the goals of public law are to prevent, mitigate or transform manipulation that does not result in identifiable individual harm.

c) Individual vs. societal harm

A major problem with manipulation by algorithmic systems ist that they may cause no or little harm to a single individual, while the harm caused to society may be profound. A case in point is recommender systems that favor sensational, hateful, or provocative content, thus leading to a polarization of society that can cause severe problems in the political and societal system.

d) Intent

Although most participants shared the view that subjective criteria such as intent can be problematic for implementation, as they are difficult to prove, they agreed that intent is an important normative criterion. Intent distinguishes legally relevant manipulation from other forms of influencing. Such use would also capture instances of contingent intent and would be based on an objectified notion of intent.

e) Autonomy

The participants agreed that capturing legally relevant instances of manipulation aims to protect autonomy and individual agency. As these are elusive concepts, interferences with autonomy and agency cannot be readily used as the sole normative criteria for assessing legally relevant instances of manipulation. Further research is needed to conceptualise autonomy and agency in such a way that they can be used as normative criteria to distinguish legally relevant cases of manipulation from acceptable forms of influencing.

f) Due process

Some participants noted that prevention of and compensating for manipulation is not only a matter of private law redress but also key to the due process rights that are core to the rule of law. This is important because manipulation is meant to ensure that people are not aware of their choices duet to a manipulative design, thus circumventing due process altogether.

4. Future research

Participants identified various topics that should be the subject of future research. These include:

4.1 Decision-making

Research should strive to determine what kind of decisions should be taken by humans alone, what decisions can be taken by algorithmic systems, and in which instances a cooperation of humans and algorithmic systems is the most promising approach.

4.2 Critical thinking

Research is needed to better understand what individuals need to know and how they should be educated and trained to ensure that they are able to interact with algorithmic systems in a meaningful and safe manner.

4.3 Transparency

Research should clarify whether individual users and/or oversight bodies such as government agencies, NGOs, journalists, and researchers should be granted varying levels of transparency of algorithmic systems. In addition, research is needed to understand how transparency can be made actionable, i.e., how the information provided can empower individuals to protect their rights and interests.

4.4 Legal enforcement

It should be investigated how collective redress can be used to enforce private law provisions on discrimination and manipulation and whether groups of individuals and/or NGOs should be able to claim damages and seek profits on behalf of individuals.

4.5 Legal framework

The development of a legal framework for the use of algorithmic systems in Switzerland should not only include the perspectives of the EU and the US. Rather, it should also include those of countries in a somewhat similar situation, such as Japan, UK, Israel, and Australia. However, this does not mean that the so-called Brussels effect and the way the US Federal Trade Commission is enforcing against potentially nefarious effects of algorithmic systems can or should be neglected.

4.6 Empirical research

Empirical research should strive to clarify what harm can be caused to individuals and society when individuals are manipulated by algorithmic systems.

* * * * *