Universität Zürich UZH

CENTER FOR INFORMATION TECHNOLOGY SOCIETY AND LAW — ITSL

Towards Principled Regulation of Automated Decision-Making (ADM)

– A Workshop Report

by Florent Thouvenin, Alfred Frueh, Teresa Rudolph, and Simon Henseler (all University of Zurich) as well as invited experts Emre Bayamlioglu (Tilburg University), Nadja Braun Binder (University of Basel), Chrisitan Katzenbach (Humboldt Institute for Internet and Society), Joshua Kroll (Naval Postgraduate School), Momin Malik (Berkman Klein Center for Internet & Society, Harvard University), Matthias Spielkamp (AlgorithmWatch).

From 12 to 14 September 2019, the aforementioned group of international experts in the fields of law, computer science and communication studies discussed issues surrounding the topic of “Automated Decision-Making” at a workshop hosted by ITSL in Lavin/Zernez, Switzerland. The workshop is part of an ITSL research project on Automated Decision-Making funded by the Hasler Stiftung and was sponsored by the Swiss National Science Foundation (SNSF).

Some of the main insights and/or ideas discussed at the workshop are the following:

1. Terminology: Automated Decision-Making

Regulators in the EU and Switzerland have enacted or drafted data protection provisions to regulate automated decision-making. At present, these provisions only pertain to decisions that are solely based on automated processing of personal data. In addition, they only include decisions that have a legal effect on individuals or that affect individuals in a similarly significant way. In principle, the General Data Protection Regulation (GDPR) prohibits automated decision-making that falls under its scope but it contains a number of exceptions (art. 22 GDPR). In contrast, the Draft Swiss Data Protection Act (Draft DPA) mainly contains information duties (art. 19 Draft DPA).

However, on closer inspection, the term automated decision-making, which aims to define the scope of application of art. 22 GDPR and art. 19 Draft DPA is neither precise nor operable for several reasons:

1.1 The term “***decision-making***” has neither a precise meaning nor does it clearly limit the scope of application. From a technical perspective, for example, everything a machine does can be understood as decision-making; most often, the decision involves producing a specific output based on a specific input by applying pre-defined rules. The term decision-making is therefore unable to draw a line between different uses of machines.

1.2 Moreover, the term “***automated***” is not well suited to capturing the phenomenon. “Automation” means solving a problem by applying a given set of rules. Automated tasks can therefore be performed by machines as well as humans.

1.3 In addition, the term “automated decision-making” can be regarded as an **oxymoron** because automation is only possible if all necessary decisions have been made and the necessary steps have been defined in the set of rules that will be applied. Therefore, automation does not capture the decision-making process, but rather moves it to an earlier stage, one where the details of particular cases may not yet be fully formed or visible.

1.4 The formulation “***solely based on*** *automated decision-making*” aims to further define the types of decisions that raise concerns and the scope of application of the regulation. However, the issues at stake do not disappear if there is some (minimal) human intervention. Accordingly, decision support systems, in which human decision-makers are supported by a system intended to improve performance, must be included as well. While we can assume that a minimal degree of machine intervention would not be a sufficient criterion for regulation, it seems impossible to define the (approximate) degree of machine intervention needed to trigger the concerns.

1.5 Given the context of data protection law, the provisions in the GDPR and the Draft DPA do not address decisions that are **not based on the use of personal data**. This is hardly satisfying, as such decisions may have similarly detrimental effects on individuals, specific groups, or society at large (*e.g.* predictive policing used to identify areas where crimes are likely to occur).

1.6 Because of these **terminological problems**, it is impossible to delineate the scope of application of art. 22 GDPR and art. 19 Draft DPA. To define the scope of application, we need to re-construct the legislators' intention when they used the term "automated decision-making" by analysing the rationale behind the regulation and the phenomena that these norms were meant to capture.

1.7 Although we acknowledge the failure of the term "automated decision-making" (ADM) to define the scope of application of a regulation, we will still use that term in the following as it is already established and allows for some sort of intuitive understanding of the phenomenon. However, the definition of the term ADM is unrelated to the established meaning of the terms "automated" and "decision-making".

2. Phenomena

2.1 The term "ADM" encompasses phenomena such as traffic lights, e-recruiting practices, facial recognition systems, spam filters, risk assessment tools, predictive policing etc. (For further examples see Automating Society, Taking Stock of Automated Decision-Making in the EU, available at www.algorithmwatch.ch/automating-society). As becomes apparent from these examples, ADM systems are used in almost all areas of life.

2.2 Some ADM systems rely on automatically found correlations, rather than a reasoned study of causality or human judgements about responsibility, for making a "decision". This can be problematic and raises concerns. For example, a bank may reject a credit application, not based on a determination of whether the applicant is solvent or intends to avoid repayment of the amount due, but because he or she lives in a district of a city in which people have failed to repay their debts. Such correlation-based decisions are far easier for banks to make, and indeed may be far more effective at decreasing their risks than trying to make reasoned judgements about specific individuals; but this effectively passes risk onto consumers, who are rewarded or punished for circumstantial (although robust) connections. As there might be an underlying causal link between where one chooses to live, or is forced to live, and failing to repay a loan, it is not where one lives that itself causes non-payment.

3. Rationale

There are different rationales for regulating ADM. The most relevant basis for regulation is the difficulty most humans have in sufficiently **understanding** what processes are executed and how those processes are related to outcomes when decisions are taken automatically. In particular, when decisions are based on correlations rather than causality, this opacity points to the need to regulate ADM.

Notwithstanding the above-expressed reservations, provisions addressing ADM may be based, *inter alia*, on the following rationales:

a) Object formula

3.1 Some argue that the rationale behind regulating ADM is that a human being should not be the object of a decision taken by a machine (object formula), as this would threaten human dignity and autonomy.

3.2 This reasoning fails to take into account that, according to the technical *status quo*, any decision requires human planning and rule-making. The object formula also misses the relevant point, which is whether human beings are treated as objects – irrespective of whether the entity treating them as an object is a human or a machine.

b) (Poor) Decision Quality

3.3 The first provisions regulating ADM were based on the presumption that automated decision-making would yield poor results, especially compared to human decisions.

3.4 This no longer holds true. In many instances, ADM systems outperform humans. If they do not, they are usually only used if they are on par with human decisions. If an ADM system yields (relatively) poor results, its use may still be attractive for other reasons such as scale, speed and lower costs. In these cases, the poor quality of the decisions may be problematic. Such problems, however, also exist with human decision-making, *e.g.* if people making decisions lack sufficient skills, information or time.

c) Transparency

3.5 ADM systems can lack transparency. Often, individuals are not aware that an ADM system has been used. As a consequence, they cannot resort to data subject rights provided for by data protection law or use alternative services that do not involve ADM.

3.6 Even if individuals know that an ADM system was used, they will (in all likelihood) have no idea how the ADM system works. This is particularly problematic with correlation-based ADM systems, which are harder to explain and understand than causality-based ADM systems.

3.7 Nevertheless, transparency is not an end in itself, but rather a means to an end. It is, for instance, an instrument to achieve autonomy and accountability.

d) Fairness

3.8 Many argue that fairness is at risk when actors use ADM. Depending on the understanding of the term, an ADM system can be considered ”unfair” if it explicitly uses protected characteristics, such as race, age, gender (*etc*.) or their proxies. According to another understanding, unfair ADM systems give rise to disparate measures of predictive performance, such as false positive and false negative rates, across groups (especially groups defined by the protected characteristics).

3.9 In addition, a decision taken by an ADM system may seem unfair if it is based on correlations that have no comprehensible connection to the issue at stake. For example, just because a place of residence is highly correlated with defaulting (*i.e*. people living in the same district have defaulted), it is not living in a place itself that causes defaulting. Perhaps there is some underlying connection between defaulting and where one chooses to or is forced to live, but using that circumstantial connection – as robust as it may be – as the basis for making judgements may be the issue of concern.

e) Ensuring a feedback mechanism

3.10 Although the word “prediction” is often used when describing ADM systems, these systems do not predict anything. Instead, they are trained to “perfect the present” and thus perpetuate the *status quo*. If the systems are not fed with new data, the reliability of the systems can be in jeopardy. The same might happen if it turns out that the correlations used do not produce convincing results. For this reason, there should be relevant incentives for users of ADM systems to implement feedback mechanisms that allow individuals being subject to ADM to raise concerns.

3.11 The implementation of feedback mechanisms should facilitate an adaptation of the criteria and an amendment of the data set used by an ADM system and ultimately lead to better results. However, simply adding additional data points created in a similar way will not be sufficient to ensure improvements.

4. Regulatory Approach

a) Scope of application

While it is clear that the scope of application of a (potential) regulation of ADM needs to be defined in a clear and concise manner, it remains unclear where to draw the line. The following criteria could be useful to define which ADM systems must be subject to a regulation:

4.1 Although no categorical distinction can be made between decisions by humans and machines, the involvement of a machine as such and the level of its involvement should be a relevant criterion. Otherwise, a given regulation would also apply to all sorts of human decisions, which was clearly not the intention of the legislator.

4.2 As correlation-based decisions raise more concerns than causality-based decisions, the use of correlations may trigger the need for regulation. However, it remains unclear whether the scope of application of a regulation should be restricted to correlation-based ADM systems or whether a regulation should also address causality-based decisions, especially because the lack of accurate data can lead to problematic results when using either approach.

4.3 Some ADM systems are (potentially) more harmful for individuals, groups or the society at large than others (*e.g*. e-recruiting systems as opposed to traffic lights). Hence, the potential impact of an ADM system on people’s life chances and social participation should be taken into consideration when carving out the scope of application of a regulation.

4.4 The number of individuals concerned by a decision taken by an ADM system might be important (*e.g.* decisions on credit applications vs. predictive policing).

4.5 A *de minimis* rule could be used to exclude trivial and clearly unproblematic ADM systems from the scope of application of an ADM regulation.

A concrete proposal for a clearly defined scope of application cannot be made at this point. For this purpose, further research is needed. Although the scope of application remains unclear, it seems possible to sketch out the normative content of a regulation of ADM systems.

b) Regulatory proposal

The rationales behind an ADM regulation do not justify a prohibition of ADM systems. Rather, the relevant concerns, such as understanding the processes carried out, whether they are fair and whether they entail feedback mechanisms can be addressed by granting far-reaching but layered transparency.

It is currently unclear which body of law is best suited to introduce such transparency requirements. To solve this issue, further research needs to be carried out. In any case, the scope of application should not be restricted to ADM systems that are based on the processing of personal data.

The proposed regulatory approach should differ for the private and the public sector.

i. Private sector

In the private sector, we suggest that the following transparency requirements apply horizontally, *i.e.* for all sectors.

4.6 In the first layer, a **basic transparency requirement** obliges any private entity applying an ADM system in an interaction with an individual to inform this individual about the use of the system. Based on this information, individuals could make an informed decision and either use a different service provider or stay with the same one. Upon request, the private entity could provide further information to the individual on the type of ADM system used and the logic involved but it would not be legally obliged to do so.

4.7 In the second layer, an **extended transparency requirement** is introduced. This obligation comes into play in specific circumstances, namely with regard to the impact of the decision on an individual, a specific group or society at large. The purpose of this extended transparency requirement is to put individuals in a position to contest ADM results affecting them. Therefore, individuals must be able to retrieve all relevant information about what happened and about what has led to the outcome of the decision. This transparency requirement is of quasi-procedural nature. It allows individuals to gather information that they can use to contest the automated decision legally, assuming such a basis exists (*e.g*. in anti-discrimination law).

4.8 Often, individuals do not use their rights. Therefore, and to accommodate concerns on a group or societal level, authorities could be vested with the power to take up cases in which the use of ADM systems has a relevant impact on a specific group or the society at large, *e.g.* if a certain group is structurally discriminated against.

ii. Public sector

In the public sector, another regulatory approach is advisable.

4.9 Any public authority using an ADM system when interacting with individuals should be required to disclose the use of this system in a dedicated public register. The register would contain the ADM system's purpose of use, an explanation of the model (logic involved) and the information on who designed the system. In addition, every individual being subject to an ADM system should be informed accordingly.

4.10 If a government agency uses ADM systems when issuing an administrative order or a court uses ADM systems to take a judicial decision, an extended transparency requirement should be introduced. Its purpose and content concur with the extended transparency requirement for the private sector (see 4.7)

iii. Sector-specific approach

4.11 The general transparency requirements in private and public law may be complemented by existing (and future) sector-specific provisions. If a sector-specific regulation exists, the applicable provisions should deal with the risks associated with the use of ADM systems. Such rules already exist in the financial market, labour, food safety, transportation safety and health sector, among others. Medical diagnostic software, for example, has to meet regulatory standards or have regulatory approval in order to be marketed; the software has to pass a test for safety and effectiveness for the intended use. In sectors that already have specific rules in place, the general transparency requirements may not provide much value at the margin; but they will not be invasive in such circumstances. In some sectors, however, the general transparency requirements will help to fill existing gaps, *e.g*. in labour relations, when an individual is subject to an ADM performance scoring system.

4.12 Against the background of existing (and future) sector-specific rules, there is no general need for a right to human intervention or a right to contest the outcome of ADM systems. It would be particularly difficult to justify the introduction of a right to contest a decision merely because this decision was taken by an ADM system instead of a human being. Instead, the aim of an ADM regulation must be to provide the individuals concerned with the information needed to take advantage of the existing legal remedies. Introducing the above-mentioned transparency requirements achieves this objective in an appropriate way.

* * *