DIGITALE | GESELLSCHAFT

Position of the Digital Society on the Regulation of Automated Decision-Making Systems

21st February 2023

Version 1.0

This publication is based on the document "Position der Digitalen Gesellschaft zur Regulierung von automatisierten Entscheidungssystemen" from the 21st of February 2022, version 1.0, and was translated from German partially automatically.

Written by the ADMS expert group:

David Sommer, Andreas Geppert, Christian Sigg, Daniel Donatsch, Erik Schönenberger, Nicole Pauli, and Tanja Klankert

Digitale Gesellschaft
CH-4000 Basel

office@digitale-gesellschaft.ch
www.digitale-gesellschaft.ch

Postkonto 61-177451-1
CH15 0900 0000 6117 7451 1

Inhalt

1. Introduction

Automated decision-making systems (ADM systems, ADMS, see glossary) have become part of our everyday lives. The algorithms (see glossary) deployed in social media and news portals decide, for example, which news we see. Private risk assessments impact whether and on what terms we receive loans and insurance benefits. The output of a predictive policing system can determine where police patrols should or should not patrol. These ADM systems thus have a significant impact on people's daily lives. They can, in some circumstances, affect individuals' development opportunities and even violate their fundamental rights. Furthermore, ADM systems can have a societal impact, for example, due to an algorithmic bias (see glossary), which may disadvantage certain individuals or groups[1].

In addition to positive benefits, ADM systems may have severe negative impacts on individuals and society. Therefore, several questions arise. For example, are the decisions made by ADM systems fair and just, or do they discriminate against certain individuals or groups of people? How can automated decisions be understood and explained? How can it be verified whether the decision taken is justified? Are there individual decisions or entire areas (e.g. court rulings) that we should not leave to algorithms, but which should only be made by humans? From these considerations, we conclude that the possible influences of ADM systems on individuals as well as on society must be subject to critical reflection. Moreover, our society must come to an agreement on which areas and in which way ADM systems should be applied. Regulation should also ensure that the benefits and risks are well balanced. That is the societal context of the proposal for a regulatory framework presented in this paper.

Transparency is essential for dealing with ADM systems. Affected individuals, eligible NGOs, as well as a supervisory authority should obtain the right to inspect applications that use such ADM systems. The proposed legal framework is based on individual assessments of the risks posed by ADM systems. By creating transparency, the assessment of the risk of an application is made possible. Depending on the individual risk, there are fewer or more restrictions and rules to be considered for the respective ADM system.

Our proposal is technology-neutral[2] and follows a "human-centered" approach: ADM systems should benefit people [3]. The regulatory framework also takes into account the fact that the risk of a system can alter over time.

1 For example, in the automated assessment of the labour market reintegration chances of unemployed persons.
2 The focus of our regulatory proposal is on the impacts and risks of ADM systems, rather than bans on specific technologies.
3 The principle of benefit

2. Scope of application

The scope of our proposal includes ADM systems that make, or at least support, decisions in a fully automated way using technical systems. We adopt the following definition from a recommendation to the City of New York by the AI Now Institute (Richardson et al. 2019, p. 20):

> An «automated decision system» is any software, system, or process that aims to automate, aid, or replace human decision-making. Automated decision systems can include both tools that analyze datasets to generate scores, predictions, classifications, or some recommended action(s) that are used by agencies to make decisions that impact human welfare[4], and the set of processes involved in implementing those tools.

ADM systems use algorithms and/or artificial intelligence techniques for decision-making and are often (but not always) data-driven. However, this does not mean that every algorithm or Big Data system shall fall within the scope of this proposal. Furthermore, the statement that almost every computer program makes decisions all the time is correct in principle, but from a regulatory perspective, it is not useful. The decisions covered by the proposed regulation must be observable as individual, discrete decisions and of a certain significance. Furthermore, they must have a potential impact on individuals and/or society to fall within the scope of this law.

If a technical system does not fall within the scope of the legal framework, a more detailed risk categorisation according to Chapter 6 is not necessary and the associated effort does not have to be made.

4 Impact on public welfare includes but is not limited to decisions that affect sensitive aspects of life such as educational opportunities, health outcomes, work performance, job opportunities, mobility, interests, behaviour, and personal autonomy.

3 Summary of the legal framework

The legal framework follows a hybrid design between a harm-based and a risk-based approach. In the first approach, sanctions are only imposed retrospectively in the event of harm, whereas in the second approach, high-risk applications are subject to certain a priori restrictions. Anyone using an ADM system must assess and categorize its risk to the health, safety and fundamental rights of individuals and society. The legal framework provides three categories for this purpose: "low risk", "high risk" and "unacceptable risk". Basically, the categories are based on the risk posed by the system to individuals as well as to society as a whole. Thus, "low risk" systems pose a low risk to individuals and none to society, while "unacceptable risk" systems pose an unacceptably high risk to individuals or society. In between are the "high risk systems". For these systems, there is a far-reaching duty of transparency and due diligence, which should enable the public to assess the risk and thus their benefit. This is because, in contrast to systems with unacceptable risk, those with a high risk are not banned.

First and foremost, we consider the impact of ADM systems on individuals. However, the widespread use of easily scalable systems may also create a risk to society[5] that is not sufficiently measurable or sanctionable only in terms of individuals. ADM systems that implement the mass distribution of individualized political advertising on social networks are an example of such a societal risk. In individual cases, the risk may be negligible with reference to individual sovereignty. However, the aggregated, total effect of such ADM systems, for example on election results, can be tremendous and thus can compromise democratic processes. Swiss jurisdiction, which has so far focused on individuals, e.e. in the case of the Data Protection Act, falls short in such cases. Our proposal addresses this shortcoming by recognizing the risk to society and providing collective remedies such as class actions and a right of action for associations.

The proposed legal framework distinguishes between ADM systems used in the private sector and those used in the performance of public duties. For both, we argue for a right of complaint or action for affected individuals, the federal ADM supervisory authority, and eligible NGOs to guarantee a meaningful risk classification and the enforcement of the related obligations.

The to-be-created federal supervisory authority for ADM Systems should be able to collect complaints and investigate the use of ADM systems in companies and government agencies on suspicion. Furthermore, the authority should be able to act as a first instance to impose administrative sanctions. As a diverse expert committee, composed of persons with sociological, technical, and legal expertise, it should be financially and personally independent and free from directives.

The details of an ADM system are usually subject to commercial secrecy. Accordingly, it is

5 https://booksummaryclub.com/weapons-of-math-destruction-book-summary/

difficult for outsiders to obtain evidence of the risk of a system. Therefore, a reversal of the burden of proof should apply. This means that a company has to prove the correct classification when responding to a complaint. For systems in fulfillment of a public mandate, we demand extensive transparency and publication of the system internals and data, in line with the demand "Public Money? Public Code!"[6] Exceptions, for example for personal data, are listed later in this article.

The federal ADM supervisory authority supports those responsible in the risk assessment of ADM systems with checklists and good practices guidelines to ensure awareness and adequate handling. Those who wrongly classify the system they operate and thus fail to meet their obligations, or operate a prohibited ADM system with unacceptable risk, are to be subject to severe and turnover-based fines. These should be of administrative nature, explicitly not aimed at punishing individual employees through criminal law, as it is usually not a matter of individual but corporate responsibility. However, these sanctions should remain the last resort.

In order not to prevent innovation, we rely on self-declaration instead of burdening companies and public administration with bureaucratic inspection processes. This allows the industries and actors in question to shape their implementation to comply with the rules themselves, within the parameters defined by the legal framework.

An increasing number of internationally significant institutions, such as the European Union or the American Association for Computing Machinery (ACM), are now addressing the need for regulation of ADM systems. We are convinced that the proposed legal framework will help to close the existing regulatory gaps.

The core aspects of the legal framework - in particular the risk categories and the transparency and due diligence obligations - are explained in detail in the following.

6 Software paid for by the public and its source code should be open and accessible to all. https://publiccode.eu/

4. The social relevance

This chapter explains why the Digital Society is campaigning for an ADM law.

Automated decision-making systems are neither objective nor neutral. This insight is central to the relevance of the present legal framework. On the one hand, systems always carry the values of their developers as well as those of society. On the other hand, they are bound by their function: Design decisions made consciously or unconsciously by developers affect the working of the system and can be negative in nature. The function of a system determines a narrow frame of action that is often not questioned.

Accordingly, automated systems can also be seen as a socio-economic mirror of a particular society. This also becomes problematic because cultural and societal values may differ around the globe, whereas technologies such as ADM systems are used across borders or globally. However, awareness of this issue is still not widespread enough in society.

The increasing support provided by ADM systems has many advantages. However, due to their usefulness, they are often perceived as objective helpers whose decisions must necessarily be correct. The reason for this is "**faith in technology**": people trust the machine that its calculation results are objective and correct.

This feeling of security and objectivity is deceptive because, for many decision-making problems, there is no optimal solution. Decisions with social implications are subject to negotiation processes and can vary internationally. In this context, delegating everyday tasks to an ADM system can lead to interaction with machines as well as their results becoming part of social reality. Sociologist Michele Willson writes in her essay: "An algorithm is given a task or a process, and the way it is used and dealt with in turn affects the things, people and processes with which it interacts - with varying consequences" (Willson 2017: 139). This feedback effect means that automated decision-making systems, as well as their (sometimes faulty or inaccurate) databases, are always changing and becoming part of the social fabric. The handing over of tasks to automated systems can also lead to a decrease in individual and collective accountability.

The effects of automated decision-making systems can be both intended and unintended. This is particularly problematic in the case of discrimination. For example, ADM systems trained on data sets reproduce the discrimination practices implicit in them. For example, recruitment systems trained on historical data would presumably still discriminate against women more often, although this is no longer tolerated. People do not make better decisions per se and are not free of prejudice. But they can reflect on them, also with the help of technology, and exchange ideas about them. **Systems lack this reflective capacity**, which is why they should not be delegated decisions that have lasting effects on society. Discriminatory effects of ADM systems can be amplified, especially in combination with technology credibility, through increased use, including across borders.

Automated decision-making systems increasingly curate the overflow of information, for

example in the form of so-called "recommender systems", "fact-checkers" or "copyright-checkers". In this way, system operators are empowered to selectively amplify political messages and positions through targeted agenda-setting. In doing so, these systems (e.g. as fact-checkers) do not necessarily have to operate on personal data. What many of these effects of automated decision-making systems have in common is that they often happen covertly and their effects are only noticed late or through further, indirect effects. The use and functioning of the systems are often unknown due to their developers and operators having no interest in disclosure.

The increasing processing speed and transmission possibilities of information allow for greater networking than was previously possible. The interaction of different ADM systems creates additional risks that are difficult to assess. The emergence of effects that reinforce each other through feedback loops (see glossary) is foreseeable and thus the resulting social risk (more on this in Appendix Chapter A). However, this **increase in complexity** is still encountered by people who face increasing difficulties in penetrating it, even with full transparency of the individual systems.

To be able to assess these effects to some extent, a **far-reaching transparency and due diligence obligation** should therefore apply. At least for important automated decision-making systems, public attention is required for a sustainable, public discourse on the norms and values underlying the metrics, the measurement or key figures that are applied, evaluated and interpreted. In summary, **the people must have final authority over ADM systems and not vice versa**.

Furthermore, many effects cannot be grasped conclusively from an individual perspective and only become visible through the accumulated observations of many affected persons. Unfortunately, however, the existing laws mostly argue from an individual case perspective. Therefore, we need methods for **collective law enforcement**, which are rarely found in Swiss law.

By focusing on impacts and risks, a **technology-neutral formulation** allows for flexibility in responding to new methods or changing uses of existing technologies. As a general goal baseline, people should benefit from the use of such systems overall. While these and other topics are outlined in this chapter for completeness only, they are discussed in academia and other work intensively.

5. An Automated Decision-Making Systems act

As a society, we are increasingly confronted with the specific effects of ADM systems. Therefore, we call for a **separate "law for automated decision making",** or at least substantial extensions of existing laws. We demand **transparency** in the use of ADM systems in order to be able to apply the laws that already exist, and effective penalties for non-compliance. We demand **explainability** of ADM systems so that humans can understand ADMS quickly and with reasonable cognitive effort. We demand constitutional control of critical ADM systems with the possibility to intervene if necessary.

We primarily see specific effects on individuals and want to give them opportunities to assert their rights. However, there are risks that are more likely to affect society as a whole, such as political influence through personalized advertising campaigns or self-reinforcing feedback effects where interlinked systems - with or without human intervention - could form their own cyclic structures. **Transparency is central, but without further measures, it is not sufficient.** The right to informational self-determination requires, in addition to knowledge of the processes, also the possibility to exercise control over them to a certain extent.

We demand **clear protection goals, namely the adherence to fundamental and human rights, the protection of mental and physical health and safety of individuals, the protection of life and development opportunities, and the protection of democratic rights and processes**. Furthermore, the public must be given the ability to effectively monitor these protection goals and intervene if necessary. Humans should have agency, i.e. they should generally be superior to the machine in their interpretation; ADM systems should enable humans to pursue ideas and goals more sophisticated, quicker and less error-prone.

The ADM law should neither inhibit innovation nor impose a disproportionate bureaucratic burden on companies and ADM supervision. We advocate a broad legal framework that introduces the needed regulation in general but allows the individual economic sectors to determine the most effective methods for implementing the protection goals themselves. Our technology-neutral and risk-based categorization, described in detail later, is **compatible with the European Union's proposed AI Act**; unlike the EU's categorization based on application areas, however, it allows for context-dependent risk classification of applications. Appendix chapter B provides an overview of other regulatory efforts, for Switzerland and internationally.

In principle, existing Swiss laws can also apply to ADM systems with some modifications. For example, the dispersion and use of personal data can be regulated by the Data Protection Act[7] . Discrimination prohibitions are the negative image of the emerging fairness discussion[8], but

7 We call for an adaptation of Art. 21 para. 1 nDSG (delete "exclusively"): "The controller shall inform the data subject of a decision which is based on automated processing and which entails a legal consequence for him or her or significantly affects him or her (automated individual decision)."

8 Fairness in relation to decision-making algorithms is about the evaluation and correction of algorithmic bias. Outputs of decision algorithms are considered "fair" if they are independent of

with a large grey area in between[9], where the bulk of real applications will be found. Labor and data protection law prohibits some surveillance practices, algorithmic and non-algorithmic, in the workplace[10]. From a formal legal point of view, a separate ADM law should be created for two reasons: Firstly, because the necessary changes in other legal texts require a common definition of terms, categorization and risks, and secondly because the ADM supervision described below is difficult to define elsewhere.

5.1 ADM supervision

The **federal ADM supervisory authority** should act as a **competence center**. It advises companies, the administration, and the public; it orchestrates long-term analyses. It collects complaints from affected parties and monitors compliance with the law and the categorization of state and private-sector ADM systems independently of instructions.

ADM supervision should be normative at all levels (federal, cantonal and communal). It can impose sanctions in the first instance in the event of violations, in parallel to complaints and legal action procedures of individuals and authorized NGOs. The competence to monitor compliance with the ADM Act and the prevention and sanctioning of risks for individuals and society is concentrated on this authority, whereby it has uniform public tasks for the entire public sector at all levels (see Chapter 8).

It supports the assessment of risks of ADM systems by providing checklists and good practices guidance to ensure awareness and adequate handling around the issue. It should be a diverse authority (composed of persons with different socio-scientific, technical and legal expertise) that acts independently of directives and with its own budget, for example, the FINMA. The details of how its supervisory activities would be implemented in the sense of these principles for the best possible protection of individuals and society remain to be determined.

5.2 IT-security

Automated systems, like other IT systems, are never completely secure and can thus be "hacked" or misused under certain circumstances. Their functions can thus potentially be

specific variables such as gender, age, etc. The exact (mathematical) formulation of fairness is still an open debate, and some definitions even contradict each other.

9 While discrimination as an offense requires a serious violation of fairness, perfect fairness is usually only possible for one specific metric and must neglect other (equally valid metrics). There is a large gray area in between.

10 The monitoring of employees is permitted to a limited extent, e.g. the recording and observance of working hours (Art. 46 ArG), data in connection with suitability for the employment relationship, etc. The limits of monitoring lie in the protection of personality (Art. 328 et seq. CO), data protection under the DPA and certain mandatory articles of the Labor Code. Systematic monitoring of workers' behavior is not permitted (Art. 26 para. 1 ArGV 3), as it may have health effects on workers. Exceptions may be permitted (Art. 26 para. 2 ArGV 3) if they are carried out for other reasons, e.g. to optimize performance or quality assurance, and only if proportionality is maintained and the risk to personality and health is kept to a minimum (case-by-case consideration) (cf. Bürgi and Nägeli 2022).

manipulated by insiders or third parties. In our view, however, measures to prevent such attacks are not in the scope of an ADM regulation, but should be addressed by a general "**IT-security law**".

Rather, the ADM Act is intended to address the specific effects of automated decision-making systems. The risk classification must not only take into account the intended use of the system, but must also consider foreseeable, possible misuse and abuse[11] ("reasonably foreseeable misuse", EU AI Act Art 9.2(c)). An example is the analysis of the communication of all employees for the purpose of improving cooperation. A foreseeable misuse of this system might be monitoring and/or evaluation of employees.

An important building block for the reliability of algorithms is the still missing application of **product liability to software** and thus also to all types of computer algorithms. It should not be possible for software producers to escape all responsibility by cleverly formulating GTCs. However, due to its generality, product liability should be part of such an IT security law and not only defined specifically for ADM systems.

11 These include the unlawful use or "hacking" of these systems, but also ADM system-specific effects, such as the extraction of sensitive training data from the models (see glossary) themselves, the unreliability of predictions for data series with which one has not tested (fragility), specially generated data series that look correct to humans but result in incorrect outputs (adversarial examples), etc.

6. Categorization of ADM systems

Our proposal intends a balance between a harm-based and a risk-based approach. In the harm-based approach, sanctions are only imposed retrospectively in the event of damage, and only after the harm has been proven. In a risk-based regulation, applications have to follow a due process at the time of their deployment. As a result of this mixed form, high-risk and high-impact applications are subject to due diligence and transparency obligations at the time of their deployment while we rely on self-declaration for less high-risk and high-impact applications. Potential breaches of duty or miscategorizations are punished a posteriori through high penalties in the context of complaints and lawsuits.

In doing so, we divide ADM systems into three categories: "low risk", "high risk", and "unacceptable risk". The classification of automated decision-making systems is based on their risk in terms of impact on individuals and society. The risk to society is based on the potential for harm and the probability of its occurrence while the risk to health, safety and fundamental rights is the primary perspective in the individual case. We elaborate on these risks in the next section. We then explain assessment criteria according to which ADM systems are to be categorized. Finally, the specific categories are explained.

6.1 Risks to "health, safety and fundamental rights" and risks to society

In the following, we address the risks to "health, safety and fundamental rights" that may result from the usage of ADM systems for individuals and risks that may affect society. For a detailed discussion of some of these points, we refer to page 43ff of the "Report of the Data Ethics Commission" of the German Federal Government.[12] [13]

For individuals as well as groups of people, the risk to "health, safety and fundamental rights" posed by ADM systems includes, in particular, the risk of violation of the right to individual self-determination and self-development: the right, in other words, to form, outwardly display and change one's individual identity, as well as to determine one's individual life goals and lifestyle, thus ensuring the development and display of one's self as an expression of human freedom.

This individual self-determination is an essential part of human dignity. The prerequisite for self-determination is the protection of privacy, the protection against misrepresentation in public, the protection against clandestine or wrong data collection and use, and the protection against a false assessment by an ADM system.

The term security relates to the collection and use of data as well as the consequence of malfunctions, potential attacks on ADM systems and manipulations with malicious intent that

12 cf. Data Ethics Commission of the Federal Government 2019

13 For a detailed analysis of the fundamental rights implications of facial recognition technology, see (FRA 2019).

may have a negative impact on the physical and mental safety and health of the individuals concerned.

We group these risks for individuals or groups under the umbrella terms "**health, security and fundamental rights**". We explicitly exclude the expansion of police and intelligence surveillance measures from the term "security".

The additional **risks to society** complement the above findings but add further dimensions that can become critical through the usage of ADM systems.

In particular, the freedom and equality of democratic decision-making, elections, and voting procedures must be protected from manipulation and radicalization by ADM systems. Automation and concentration of power by media intermediaries with a gatekeeper function pose a considerable threat to democratic decision-making and democracy. "Chilling effects" through surveillance (cf. Penney 2016) and their negative impact on the exercise of fundamental rights (cf. Assion 2014) are already a reality.

Education plays an outstanding role in securing a free democratic basic order, as it influences in many ways the critical participation of citizens in shaping society, which is constitutive for a democracy, the understanding and assessment of socially relevant contexts and developments, and thus ultimately also the trust in a value-based future that can be shaped. Justice and the perception of justice are also related to the capability of participation in social processes. Accordingly, the influence of automated systems in these areas is far-reaching. The possible participation of the broad public in automation mechanisms is essential to strengthen participatory processes and solidarity and to prevent the systematic exclusion of large population groups.

Furthermore, the risk of reinforcing feedback loops, which has already been mentioned, should also be considered here: as the complexity and interconnectedness of systems increases, so does the risk that dynamic (information) loops arise, the effects of which a person is exposed to but does not directly participate in the responsible mechanics.

6.2 Assessment criteria

We propose a categorization by means of the risks now explained. For the categorization, we adopt and supplement some of the guidelines from the AI Act of the EU Commission (Art 7.2). When assigning an ADM system to a risk category, the following aspects should be considered:

- What is the purpose and scope of the ADM system?
- To what extent will the ADM system (probably) be used (selectively or in a wide scope)?
- To what extent is it known that the usage of the ADM system has already caused harm to health, impaired safety or caused adverse impacts on fundamental rights? Based on reports or documented allegations that should be provided to the responsible authorities, is there cause for significant concern about the occurrence of such harm,

impairment or adverse effects?
- What is the potential magnitude of such damage, harm or adverse effects, particularly in terms of their intensity or their capacity to affect a large number of people?
- To what extent are potentially harmed or impaired persons dependent on the outcome produced by an ADM system, and to what extent are they reliant on the ADM system because, in particular, for practical or legal reasons, it is reasonably difficult to avoid using the ADM system?
- To what extent are potentially harmed or impaired persons vulnerable to the user of an ADM system, in particular due to an imbalance in terms of power position, knowledge, economic or social circumstances or age?
- To what extent and how easily can the outcome produced by an ADM system be reverted? Results that affect the health or safety of people cannot be considered easily reversible.
- Can destructive or self-reinforcing feedback loops (see appendix) arise? What measures are taken against them?

When categorizing an ADM system, i.e. assigning a system to a risk category, all these aspects must be considered in combination. In particular, avoidability and revisability are important aspects here. If an individual can avoid being exposed to an ADM system under the assumption of average knowledge and normal circumstances and without exceptional disadvantages, then this system falls into a lower category than if an individual is dependent on this system. If the effect of an automated decision can be (easily) reversed (or compensated), this system also falls into a lower category. The prerequisite for this is that individuals not only have knowledge of the automated decision itself but also of the possibility of objection and reversal; and that this reversal can be demanded with a normal level of knowledge and can take place promptly without accepting disadvantages.

6.3 The categories

If a specific technical system falls under the scope of this law (i.e. if it is an ADM system according to chapter 2), it shall be classified into one of the following three categories: "**low risk**", "**high risk**", and "**unacceptable risk**". The due diligence and transparency obligations listed below only apply to "high risk" systems.

In this classification, the risks according to chapter 6.1 are taken into account, and the assessment criteria according to chapter 6.2 are applied. The assessment of whether it is an ADM system and what risk is associated with it is carried out as self-declaration by the entity deploying the ADM system. This is to keep the bureaucratic effort as low as possible. In case of false or too low self-declaration, depending on the culpability, high and turnover-dependent administrative sanctions are possible. The assessment will, therefore, ultimately fall to the courts.

This risk-based categorization is compatible with the application-based formulation of the European Union AI Act (EU AI Act). However, unlike the EU AI Act, we do not fundamentally

ban applications, but consider them in light of the circumstances. For example, algorithmic emotion recognition may be banned in recruitment interviews; an art exposition may use it because the risk to society and the individual is low in this case. The risk of automated decision-making systems, and thus their assessment, may also change over time as technology and society evolve and interact with other systems. The proposed categorization scheme can reflect these developments.

Systems that fall into the lowest category ("no/low risk")

- pose a low risk to society, and
- for individuals
    - pose no (or only a minor) risk to health, safety or fundamental rights.

Systems in this category are thus characterized by the fact that they are unlikely to have any particular negative impact on individuals or society. If moderate harm or interference with fundamental rights is possible while the system can be avoided easily and without major special knowledge; or if harmful effects can be reversed easily, a system can still be placed in this category. Systems whose decisions can have a damaging effect on the health and/or safety of persons cannot be placed in this category.

Systems that fall into the medium category ("high risk")

- represent a high risk for the company or
- for individuals
    - pose a high risk to health, safety or fundamental rights.

The systems in this category are characterized by the fact that their (positive) benefit is outweighed by a significant potential drawback. The potential for harm is still acceptable (or mitigable), or else such a system would be placed in the next higher category. Systems in this category are typically deployed widely (not just specifically) and leave individuals with no way to escape automated decision-making. In this category, any harm caused by decisions cannot realistically be undone or compensated for.

Decision-making systems that recommend content (be it news as in newsfeed algorithms, videos in recommendation algorithms, or general content as in search engines) belong to the "high risk" category for several reasons: they affect large groups of users (potentially the whole of society), they influence the perceptions of their consumers, and they can demonstrably contribute to radicalization (cf. Tufekci 2018, Frenkel and Kang 2021).

Individual decisions in the social welfare system (e.g. eligibility assessments) are high-risk for several reasons: they usually affect vulnerable people, cannot be avoided/circumvented, and it is usually not possible for those affected to obtain corrections of wrong decisions without complicated and expensive legal action. Decisions leading to the recruitment or selection of people in job application procedures also have a high risk, as they cannot be circumvented by the people affected, but impact the life and development chances of these people.

In addition, there are systems with the potential to have irreversible and severe effects on

individuals, e.g. in medical diagnostics, and would thus be categorized as "unacceptable". As long as these systems are used as support and the final decision is made by a professional, downgrading to "high risk" is reasonable. However, the transition from support to recommendation systems to unquestioned decision-making is fluid, especially when combined with the aforementioned faith in technology, which could mutate initial support systems into de facto decision-makers.

Systems that belong to the highest category ("Unacceptable risk")

- pose an unacceptable risk to society as a whole; or
- for individuals
    - unacceptable risk to health, safety or fundamental rights; or
    - represent irreversible and serious effects.

This category, therefore, includes systems whose potential damage is of such severity that it must not be risked. For many systems in this category, the harm is also known and documented (and no longer potential, so that, strictly speaking, one can no longer consider them as a risk). The decisions in this category are neither circumventable (e.g. biometric mass surveillance) nor revisable; they are also often not reviewable (e.g. automated/supportive asylum, probation or court decisions). The expected or proven harm to individuals and society in this category is so high that the risks cannot be accepted or mitigated, and the deployment of such systems is prohibited.

Examples of unacceptable effects on society are the aforementioned biometric mass surveillance (incl. facial recognition[14]), which not only represents a massive encroachment on fundamental rights such as human dignity, autonomy and privacy but also has the aforementioned "chilling effects" (cf. Assion 2014 and Penney 2016) on democratic processes and thus on society. Another example is the automated assessment of behavior (social scoring), which primarily affects individuals but can have far-reaching (and, moreover, not democratically legitimized) effects on society through its control and shaping effects.

For individuals, in addition to asylum, parole or court decisions, we also see unacceptable effects in monitoring employees, pupils and students. Extensive automated assessment in the workplace and subsequent dismissal or optimization decisions can unacceptably damage the physical health of workers.[15]

14 for example https://gesichtserkennung-stoppen.ch, https://reclaimyourface.eu

15 For these reasons, there have already been calls to add surveillance and automated management in work and educational contexts to the list of applications to be banned (cf. EDRi 2021); on the criticism of automated emotion recognition techniques (cf. Crawford 2021)

7. Due diligence and transparency obligations

In principle, the entity using the ADM system from the point of view of the affected parties (for example, the service provider) is responsible for its functionality and correct classification into the risk categories mentioned above. It should be possible to pass on certain business risks to the manufacturers of components or systems under civil law. However, it should be prevented (via the product liability of the separate IT Security Act mentioned above) that the manufacturers can release themselves from any responsibility by means of the general trading conditions (GTC), as is currently common practice in software usage contracts.

We consider a certification obligation to be meaningful only in special and private-sector areas of application with a specific and easily standardized purpose[16] such as medical products, e.g. an automatic defibrillator (AED). Otherwise, there is a danger that accountability will be outsourced to certificate issuers on a large scale.

The following transparency obligations only apply to ADM systems in the "high risk" category. "Unacceptable" systems may not be used, period. A false or too-low declaration will be punished with severe penalties. The level of transparency obligations should enable the assessment of individual systems with regard to their risks and provide sufficient information to assess the impact of the entire ADM ecosystem. We, therefore, recommend standardised transparency reporting formats.

We distinguish between obligations for systems used in the **private sector** and those used **in fulfillment of a public mandate**. In general, all systems (private and public) with the categorization "high risk" are **subject to a labeling and notification obligation** which

1. indicates that an ADM system is being used[17] ,
2. a short abstract on the purpose of the system and concrete possible outputs, and
3. Information about the data origin and explanations about the specific features used by the ADM system (see glossary) and what they represent.

The information on the origin of the data should ensure that the data quality is sufficiently high and that the **interlinking** of several ADM systems (possibly from different manufacturers) becomes sufficiently visible. Furthermore, we demand a periodic and continuous review of the risk (i.e. the classification of the category) and the documentation regarding the transparency obligations, especially for systems that continue to learn independently.[18]

Data quality is about ensuring that the data used either matches reality,[19] meaning that

16 This means a clearly verifiable functioning of the independently deciding system.
17 with the amendment of Art. 21 para. 1 nDSG (delete "exclusively")
18 These are systems that continuously adapt their functioning based on new inputs. This mode of operation leads to a variety of problems, such as systems unlearning previously attested guarantees such as "fairness" or that they can be deliberately fed manipulated data in a way that is difficult to attest in order to change their mode of operation to one's own advantage.
19 This means that they are statistically representative (in terms of the system's objective and field of

systems based on it function as error-free as possible, or that it has only been changed in intended and generally useful aspects, for example, to prevent discrimination.

With regard to the **data origin**, the right to informational self-determination must be obliged; this also applies to data from abroad. Data must originate from ethically justifiable sources; for example, the usage of illegally obtained data is not allowed as a matter of principle[20].

Furthermore, it should be specifically stated when the output of other ADM systems is used. This should create transparency for the **interlinking** of such systems, which create their own risks due to the foreseeable increasing complexity of their interaction, their (opaque) information flow, and the resulting feedback loops.

In the following, we specify the transparency obligations for the private sector and public contexts.

7.1 Private sector

For entities using systems within the private sector context, we require information on the origin of all data used, as well as on the quality and completeness with regard to the purpose of the ADM system. That includes all data used to set up, train, validate, predict, etc the system. It also includes documentation on the purpose of the system and meaningful information on what features are used as input to assess the impact on individuals and society or the risk to health, safety or fundamental rights of individuals or society.

A formal law may provide for exceptions to the transparency obligations as well as liability for products that can be standardized, weighing up the risks and benefits, if it simultaneously creates a certification body that fulfills the quality requirements mentioned above at least equivalently (for example, in the case of medical diagnostics).

7.2 In fulfillment of a public mandate

In addition to the same transparency obligations as for systems in the private sector (information on data origin and quality, features and purpose), we call for the disclosure of coefficients (see glossary) in standardized format[21] as follows:

- For ADM systems based on non-personal data, the data should be made available as OpenData together with the coefficients as far as possible.
- For ADM systems based on personal data or on non-personal data that may not be published, the following applies: they must either a) be trained on synthesized data (synthetic data set, see glossary) and these must be published together with the

application), accurate, complete and as free of contradictions as possible, and follow known semantics.

20 In other words, the use is only justified from an ethical point of view under certain circumstances after weighing up the advantages and disadvantages (cf. Imhasly 2021).

21 This makes the automatic evaluation easier.

coefficients, or b) neither the data nor the coefficients are published if (in exceptional cases) the generation of synthesized data and the use of corresponding ADM systems involves disproportionate effort or if personal data or non-personal data that may not be published can be derived again from their coefficients. However, in this case, ADM supervision and authorized NGOs must be given access to verify the impact on individuals and society or the risk to health, safety or fundamental rights of the individual or society.

The general disclosure obligation demanded here corresponds with the demand "Public Money? Public Code!" - the demand for source code publication of software financed by public money. This means that the solutions can also be used and further developed by other authorities or the public.

8. Control, measures, and sanctions

Violations of the due diligence and transparency obligations listed above must be sanctioned effectively. Here, too, we distinguish between private-sector and public-sector use. In both cases, compliance with the due diligence and transparency obligations is monitored on the one hand by individuals and on the other hand by authorized associations (NGOs), which can lodge a complaint or file a lawsuit in the event of damage. Associations should be entitled to file complaints if they are active throughout Switzerland and have enshrined the purpose in their statutes. The possibilities for control, measures and sanctions, and the channels of appeal are to be designed in such a way that those affected can be guaranteed the best possible protection; where necessary, these are also to be augmented or redesigned. This also includes the revision and improvement of collective redress mechanisms.

The ADM supervisory authority should be able to investigate violations of the law ex officio and issue formal orders. It can demand inspection and impose sanctions. To ensure the best possible protection for those affected, both intentional and negligent actions should be punishable. We expect questionable systems to become known quickly in order to attract the attention of civil society and thus of legitimate associations or ADM supervision.

Incorrect categorization of ADM systems, e.g. as "low risk" instead of the correct "high risk", and the associated violations of due diligence and transparency obligations are prevented by making the expected sanctions high enough. For this reason, we consider the proposed self-declaration obligation to avoid bureaucratic processes and to relieve the companies to be sufficient. We expect the companies to independently implement rules analogous to data protection, for example by setting up internal reporting offices in case of suspected violations or false declarations.

8.1 Private sector

Proving individual culpability does not seem to be effective, as violations of the above-mentioned due diligence and transparency obligations are usually organizational culpability. The ADM supervision authority should, therefore, punish companies by administrative sanctions and not sanction individuals by means of criminal law. That also eliminates the otherwise threatening "shifting" of blame onto "scapegoats". Furthermore, the range of penalties must depend on turnover, so that large companies cannot get off lightly in comparison. The penalties must be sufficiently high so that violations of due diligence and transparency obligations are not perceived as an everyday business risk and thus "budgeted for".

Too low an assessment of the category of the ADM system by the responsible person is punishable.

The ADM supervisory authority has the following instruments at its disposal: it collects complaints, it can demand inspection, and it can issue sanctions and orders, as is the practice

at FINMA, for example. The final assessment is the responsibility of the courts.

In the case of suspected inadequacy, we see the following avenues of recovery:

1. Affected individuals may bring actions against private sector entities and appeal against ADM supervisory authority decisions if the ADM supervisory authority decisions are deemed inadequate. Class actions and complaints should also be explicitly possible. The appeal procedures are to be adapted in this sense.
2. Entitled associations (nationwide and with a suitable purpose according to the statutes) should be able to file a lawsuit against private entities and complaints against ADM supervisory decisions without being personally affected (right of association to file a complaint or lawsuit). In view of the high costs of litigation and as a regulatory element, the association in question can receive a portion of the sanctioned amount as compensation for its expenses.

In the case of a threatened conviction, the concealment interest of the accused leads to a strong imbalance of power. We, therefore, call for the reversal of the burden of proof so that accused entities must sufficiently prove that they have not violated categorization requirements, due diligence or transparency obligations if a court recognizes the complaint or lawsuit as admissible.

8.2 In fulfillment of a public mandate

In principle, the same controls, measures and sanctions should be possible as against private entities. How the relationship between ADM supervision authority and the entities in fulfillment of a public mandate at the cantonal and communal level is to be structured in detail remains to be clarified.

There should be possibilities for both individuals and associations (analogous to Chapter 8.1 Private sector) to take action against risks emanating from ADM systems as well as against the results of such systems. In order to avoid possible conflicts of competence, for example, if entities with a public mandate at the communal or cantonal level are affected, the ADM supervisory authority could always have party rights in the cantonal proceedings.

9 Suggestions for the future

Automated decision-making systems will take over more and more tasks, work, and functions in the future, which may result in new consequences, opportunities, challenges and problems. In the following, we want to address various points for which regulatory intervention could become necessary.

The first point concerns the **question of power**: who creates, determines and controls the systems, algorithms and metrics used? The relevant people and organizations have a strong influence on the perception and possibilities of our natural and social environment. Accordingly, it is essential to observe closely how these dependencies develop.

The second concerns the **linking and chaining** of wide-ranging automated systems: In the near future, the output of one system may be partly the input of the other system, which in turn may have an influence on the first system. That can lead, especially with more than just two systems, to complex and multi-layered feedback effects (see appendix) and thus risks that are difficult to assess. The foreseeable, partial intransparency of the interlinked systems and the associated unpredictability of these effects will make it necessary to address them. Potential solutions would be a clear modularization of the systems, so that the internal action of the systems can be reduced to a simple abstraction, and that this is sufficient to estimate the consequences of the feedback. It is also conceivable to prohibit coupling for systems above a certain cluster size or if certain security or purpose limitation criteria are no longer fulfilled.

In certain circles, there is a vision that all social and personal problems can be solved with more data and better algorithms if only they are allowed. This worldview tries to squeeze the complete reality into a construct of formulas and numbers. Based on our insight that there is no absolute objectivity and, therefore, that all metrics, measurements and figures and their interpretation are subject to social negotiation processes, we see this path as misleading. We advise general **data reduction and data economy as a basic principle**, since this, as with data protection, reduces the arising problems at its source.

Furthermore, a **dependence on ADM systems** is foreseeable. The usage of automation to facilitate and reduce workloads makes it possible to accomplish increasingly complex tasks in less time. However, we should be aware of what a failure of these automated systems would mean for us, what impact it would have, and what risks would be associated with it, and prepare quickly implementable emergency strategies as an immediate counter-measure. Increasing interconnectivity and dependence on single resources, such as the role of the internet, also increases the risk that many systems could fail simultaneously. It might make sense to consider (regulatory) measures that establish redundant systems.

At the same time, one can also anticipate a potential **loss of human competence and responsibility**. The handing over of tasks to automated systems, the reliance on correct execution, and the associated habituation effects could lead to an unlearning of competencies that would be needed without corresponding systems and to a lower individual or collective

responsibility. Corresponding effects could possibly only become apparent over the course of several generations, for example when certain skills are no longer passed on.

Simple jobs without long training requirements will be becoming increasingly scarce. That can have considerable social consequences, also in Switzerland, since social status and work, among other things, are closely linked. We need to reflect on our understanding of social esteem and, thus, on the accumulation and distribution of wealth **such that the benefit of automation enriches the life of everyone**.

Finally, we would like to emphasize the importance of continuous **technology impact assessments**, such as those carried out by TA-Swiss on behalf of the Swiss Confederation, among other organizations. In general, technology impact assessments aim to produce a systematic analysis and evaluation of the effects and consequences of technologies in all sub-areas of the natural and social environment that are obviously affected.

Appendix

A. Feedback loops

In the ADM environment, feedback loops describe the effect of the results of AMD systems on their input or on the input of similarly acting systems. These effects are manifold and can span multiple ADM systems and their interaction with people's (reactive) behavior. Because of these complex interactions, these feedback loops are difficult to detect and often only become indirectly visible after some time. They become particularly problematic when harmful dynamics become self-reinforcing.

Ensign et al. (2018) describe a class of such feedback loops using the example of predictive policing systems (cf. Ensign 2018). They prove (mathematically and empirically) that the prediction of impending crimes in certain regions led to more police presence in these regions. As a result, a proportionally higher number of crimes were detected there than in regions with less police presence (because more eyes were watching), which then led to even more police presence in the pre-stressed operational regions. This dynamic is self-reinforcing.

The basic problem with this type of "runaway" feedback loop is that only one type of result finds its way back into the systems (the detected crimes) and others are hidden (the non-detected crimes in other regions). This leads to statistical distortions of the initial data situation, which forms the training basis for this predictive policing algorithm. The same applies to systems for the selection of applicants for a job, where often only those persons who were part of the training data that already received a job offer, neglecting the dismissed majority of applicants. Another example is the numerical rating of universities in the USA, which has led to an expensive "arms race" in terms of the services offered by the university and has resulted in a sharp increase in semester fees due to the resulting increased capital requirements (cf. O'Neil 2016).

Another type of feedback loop is illustrated by Cathy O'Neil (2016) using the IMPACT teacher evaluation system. IMPACT sought to assess teachers' teaching performance as the difference between the predicted 'natural' development of students based on their background and past performance, and tests written by their students, in order to make regular teacher dismissals on this basis. This in itself is problematic, as the complex social environment of adolescent students cannot be expressed in just a few numbers, and as 20 or 30 students in a class are clearly too few to obtain statistically robust results for such models. Therefore, the output of this algorithm to the same teacher over several years did not correlate strongly, which contradicted the intended function of the algorithm (the objective evaluation of teachers). In addition, some teachers began to manipulate their students and the tests in ways that were favorable to them and their future as tenured teachers, to the detriment of those who did not. This kind of destructive social feedback loop could not be corrected by the algorithm. The project was abandoned.

Mainly due to their limited scope of application and effect, the examples of crime prediction or

teacher evaluation mentioned above were corrected by human insight, not automated. That allowed to apply appropriate boundary conditions or to abandon the projects completely. Comparable effects, however, can also occur over much more complicated linkages of systems and society, where correction by individual parties is no longer possible or an obvious solution is not apparent. Predictive policing, for example, is an online learning system, but the problem also arises when new ADM systems are regularly created based on newly collected data sets. In this case, the feedback loop is constructed via the social bias in the training data (see glossary) over multiple systems. It is, therefore, not sufficient to consider online learning systems only for this effect.

B. Regulatory proposals for ADMS, Artificial Intelligence, and algorithms

In the course of digitization, (data-driven) computer systems have spread almost unregulated; apart from mostly European data protection laws. Depending on the aspect to be emphasized, these systems are called "Automated Decision Systems", "Algorithmic Systems", "Artificial Intelligence" or "Big Data Systems". Despite all the differences, they have in common that very large amounts of data are processed and analyzed with the aim of automating decisions and/or processes. In recent years, a broad consensus has emerged that these systems must be regulated in order to avoid the strongest negative risks and effects. The demand for regulation comes not only from civil society, but also from politics, business, and research.

For example, the ACM (Association for Computing Machinery) already developed a position on transparency and accountability of algorithms in 2017 (cf. ACM 2017). The ACM is the professional association of US computer scientists and thus the organization which consists of the people who drive the research, development, and use of ADM systems. Many of the statements and principles in this position, for example regarding transparency and data, are also reflected in our proposal.

There are also repeated calls from the business community for politicians to regulate such systems. Particularly impressive is the statement made by Microsoft President Brad Smith in 2018, in which he emphasizes that facial recognition must be regulated due to its dystopian potential and its dangers for democracy (cf. Smith 2018).

Previous approaches and proposals (cf. DEK 2019, EU AI Act 2021) follow a risk-based approach in which an attempt is made - as our proposal does as well - to divide ADM systems into categories based on their intrinsic risk and to define stricter rules for the categories as the risk increases. The number of categories varies between the proposals. However, there is always the lowest category (low-risk) with very few rules or requirements, as well as a category of ADM systems classified as very risky, whose use is prohibited. In the proposal of the Data Ethics Commission, for example, the risk pyramid was developed.

The EU Commission's proposal "AI Act" (cf. EU AI Act 2021), which was published in April 2021, also follows this risk-based approach. In contrast to our proposal, the EU AI Act contains concrete lists of prohibited (such as "remote post biometric identification") and high-risk applications. Since its publication, the AI Act has been the subject of intense debate.

While there is broad agreement on the necessity and relevance of this proposal as well as its general, risk-based approach, there is also detailed criticism from civil society (for example, the Digital Society with a large alliance of the EDRi network demands, among other things, a broader version of the prohibited and high-risk categories or deletion of the exceptions, especially in the area of biometric identification, cf. EDRi 2021). Similar criticism also comes from consumer protection organizations (cf. VZBV 2021). France already introduced a publication obligation in 2016 for ADM systems of public authorities that make automated administrative decisions for individuals (cf. Journal officiel 2016).

For Switzerland, there is already a position paper on the regulation of AI (cf. Thouvenin et al. 2021). Like our proposal, this paper also emphasizes the need for regulation. In contrast to the Digital Society's proposal, however, it leaves it largely open to how this regulation could look.

The AI Now Institute has compiled a whole series of governmental "use cases" for the city of New York (cf. AI Now Institute 2018), many of which are also transferable to Switzerland. The report also contains further references and motivating examples.

References

References regarding regulatory proposals of ADM systems in the European and intercontinental context

- ACM U.S. Public Policy Council (2017): Statement on Algorithmic Transparency and Accountability
- CAHAI - Ad-hoc Committee on Artificial Intelligence of the Council of Europe (2021): A legal framework for AI system (https://edoc.coe.int/en/artificial-intelligence/9648-a-legal-framework-for-ai-systems.html)
- Datenethikkommission der Bundesregierung (2019): Gutachten der Datenethikkommission der Bundesregierung. Berlin: Bundesministerium des Innern, für Bau und Heimat. Accessed at https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.html (Accessed on 16.12.2019).
- EU AI Act (2021): Vorschlag für eine Verordnung des europäischen Parlaments und des Rates COM/2021/206 vom 21. April 2021 zur Festlegung harmonisierter Vorschriften für Künstliche Intelligent (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union
- EU Digital Service Act (2020): Vorschlag für eine Verordnung des europäischen Parlaments und des Rates COM/2020/825 vom 15. Dezember 2020 über einen Binnenmarkt für digitale Dienste (Gesetz über digitale Dienste) und zur Änderung der Richtlinie
- European Commission adoption consultation (2021): Artificial Intelligence Act. EDRi (https://edri.org/wp-content/uploads/2021/08/European-Digital-Rights-EDRi-submission-to-European-Commission-adoption-consultation-on-the-Artificial-Intelligence-Act-August-2021.pdf)
- Europan Union Agency for Fundamental Rights (2019): Facial recognition technology: fundamental rights considerations in the context of law enforcement. European Union Agency for Fundamental Rights, November 2019 (https://fra.europa.eu/en/news/2019/facial-recognition-technology-fundamental-rights-considerations-law-enforcement), (Accessed on 08.02.2022)
- Florent Thouvenin, et al (2021): Ein Rechtsrahmen für Künstliche Intelligenz, Digital Society Initiative Universität Zürich
- Journal officiel (2016): Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, Journal officiel de la Republique francaise du 8 octobre 2016
- Richardson, Rashida et al. (2019): Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force. AI Now Institute ainowinstitute.org/ads-shadowreport-2019.html
- Verbraucherzentrale Bundesverband (2021): Artificial Intelligence needs Real-World Regulation, https://www.vzbv.de/sites/default/files/2021-08/21-08-

03_vzbv_Position_Paper_AIA_ENG.pdf

Other references

- AI Now Insititute (2018): Automated Decision Systems - Examples of Government Use Cases, https://ainowinstitute.org/nycadschart.pdf (Accessed on 08.02.2022)
- Assion, Simon (2014): Überwachung und Chilling Effect, in: «Überwachung und Recht», Tagungsband zur Telemediucs Sommerkonferenz 2014, epubli GmbH, Berlin
- Bürgi, Urs (2022): «Arbeitnehmerüberwachung», in: Law Media, 2022, https://www.arbeits-recht.ch/fuersorgepflicht-datenschutz/arbeitnehmerueberwachung (Accessed on 25.01.2022), Bürgi Nägeli Rechtsanwälte
- Crawford, Kate (2021): Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven, MA
- Ensign, Danielle, et al. (2018): Runaway Feedback Loops in Predictive Policing. FAT 2018: 160-171
- Fanta, Alexander (2020): Datenschutzbehörde stoppt Jobcenter-Algorithmus. Netzpolitik.org 21.8.2020 (Accessed on 7.8.2022)
- Frenkel, Sheera and Kang, Cecilia (2021): An Ugly Truth. Harper Collins Publishers
- Gesichtserkennung-stoppen.ch (2021): https://www.gesichtserkennung-stoppen.ch/ (Accessed on 14.02.2022)
- Imhasly, Patrick (2021): Artikel Forschung an Raubgut, in: NZZ am Sonntag on 19.09.2021
- O'Neil, Cathy (2016): Weapons of Math Destruction. New York: Crown
- Penney, Jonathon (2016): Chilling Effects: Online Surveillance and Wikipedia Use, in: Berkeley Technology Law Journal, Vol. 31, No. 1, p. 117, 2016, Available at SSRN: https://ssrn.com/abstract=2769645
- Public Code, Public Money: https://publiccode.eu/ (Accessed on 14.02.2022)
- Reclaim your Face (2021): https://reclaimyourface.eu/ (Accessed on 14.02.2022)
- Smith, Brad (2018): Facial recognition technology: The need for public regulation and corporate responsibility, In: blogs.microsoft.com (https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/) (Accessed on 08.02.2022)
- Tufekci, Zeynep (2018): YouTube, the Great Radicalizer, in: The New York Times, 10.3.2018, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html (Accessed on 8.2.2022)
- Willson, Michele (2017): Algorithms (and the) everyday, in: Information, Communication & Society, 20:1, 137-150, DOI: 10.1080/1369118X.2016.1200645

Glossary

- **ADM systems, ADMS**: See automated decision-making systems.
- **Algorithm**: An algorithm is an unambiguous and step-by-step procedure for solving a problem or a class of problems, which arrives at a solution after a finite number of steps. People can also execute algorithms using pen and paper.
- **Automated Decision-Making Systems**: Any software, system or process that aims to automate, support or replace human decision-making. Automated decision-making systems can consist of tools for analyzing data sets that produce evaluations, predictions, classifications or recommendations for action, or they can be understood as the processes that implement such tools (according to AI Now, Richardson et al. 2019, p. 20). They can be used to make decisions that have an impact on the well-being of individuals and society as a whole. This well-being includes (but is not limited to) decisions about sensitive areas of life, such as educational opportunities, health outcomes, work performance, job opportunities, mobility, interests, behavior and personal autonomy.
- **Bias**: Algorithmic bias occurs when a computer system reflects the implicit values of the people involved in coding, collecting, selecting or using data to train the algorithm.
- **Coefficients**: Model coefficients are specific numbers that are used in calculating the output of the model, given the input data series. These are, for example, the weights in neural network models or the discrimination limits in decision tree algorithms.
- **Data**: Also data set. A collection of data series used for setting up, training, validating, predicting (and so forth) of ADM systems.
- **Data series**: A collection of numbers, text, pictures, graphs, etc (for computers, these are all numbers) that relate to an individual, a specific event or a measured circumstance.
- **Features:** Features are attributes of the data series or derived data attributes, used as input data series for the decision algorithm (the model). These can be, for example, age, postcode, mineral water preference, but also derived meta-variables such as nutritional health.
- **Feedback loop**: In the ADM context, feedback loops describe the effect of the results of ADM systems on their input or on the input of similarly acting systems. For detailed explanations, see Appendix Chapter A.
- **Model**: A decision algorithm that can recognize certain types of patterns and relationships in input data series. In the process, the input data series are combined with the coefficients of the model (according to its architecture) to compute the output data.
- **Synthetic data set**: Artificially generated data series that correspond to real data series in all essential characteristics. The use of synthetic data avoids data protection problems when using sensitive data such as personal data. Synthetic data sets are generated artificially and their individual data series cannot be assigned to any real

person or object. But they can correctly map the properties that a specific algorithm wants to predict on them, so that algorithms trained on these synthetic data can also correctly infer the corresponding property on real data series. Simply put, synthetic datasets have the same relevant properties as real datasets, so one can train algorithms on them that work on synthetic as well as real datasets. However, as soon as properties are to be derived from synthetic data sets that were not taken into account when they were created, this can fail.

- **Training data**: A collection of data series used for the development or training of ADM systems.
- **Validation data**: A collection of data series (typically independent of the training data) to evaluate the accuracy of a trained ADM system.