September 2024
Version 2.0

DIGITAL SOCIETY

POSITION PAPER

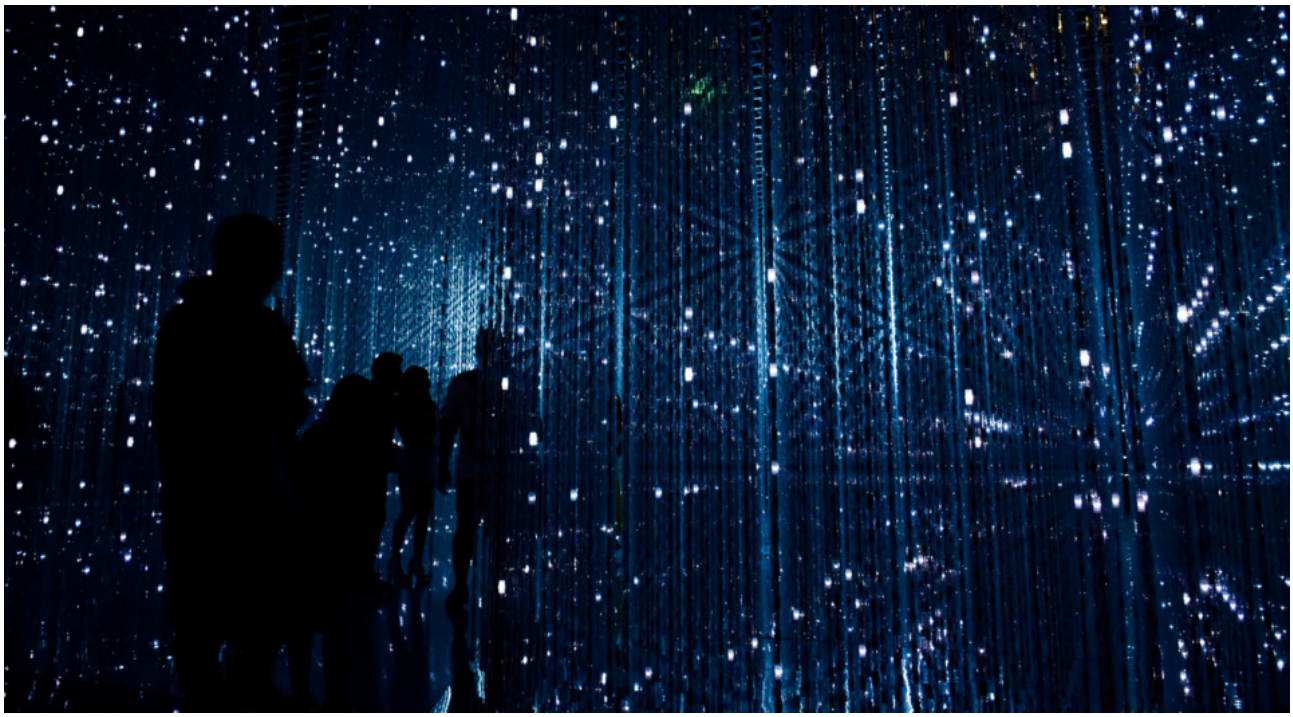
ON THE REGULATION OF AUTOMATED DECISION-MAKING SYSTEMS

Contributors Version 2.0:

David Sommer, Lars Lünenburger, Erik Schönberger, Andreas Geppert, Luka Nenadic, Christian Sigg, Tanja Klankert, Peter Fröhlich, David Caspar, Thomas Mandelz, and Jan Zwicky

Contributors Version 1.0:

David Sommer, Andreas Geppert, Christian Sigg, David Caspar, Erik Schönenberger, Jan Zwicky, Lars Lünenburger, Luka Nenadic, Peter Fröhlich, Tanja Klankert, and Thomas Mandelz



DIGITALE GESELLSCHAFT / CH-4000 BASEL
OFFICE@DIGITALE-GESELLSCHAFT.CH / DIGITALE-GESELLSCHAFT.CH

Executive Summary

Automated decision-making systems (ADM systems) offer great social and economic potential. However, they also pose significant risks to individuals and society, against which the current legislation in Switzerland offers insufficient protection. The Digital Society is therefore presenting its proposal for a legal framework to regulate ADM systems, which is based on the following five pillars:

- When using ADM systems, the protection of individuals and society must be ensured. The fundamental **protection goals** with regard to individuals are compliance with fundamental and human rights. When many individuals, society-wide processes or democratic organizations are affected, we speak of protection goals in relation to society. The protection goals are based on the data protection concept of the Digital Society.
- ADM systems are divided into **three risk categories**: ADM systems have a *'low risk'* if they are unlikely to have any significant negative impact on individuals or society. An ADM system with *'high risk'* – for example, when used in sensitive areas such as social services – carries a significant potential for harm to individuals or society. Finally, ADM systems with an *'unacceptable risk'*, such as biometric mass surveillance, will be banned outright. In order to avoid unnecessary bureaucracy, the classification is based on *self-declarations*. The accuracy of these declarations is ensured by subsequent sanctions in the event of false declarations.
- In order not to inhibit innovation, companies should not be overburdened with excessive ex ante obligations. However, this leeway should not be abused as a free pass for irresponsible use of ADM systems. It is essential that companies and public authorities fully accept responsibility for any damage caused by the use of ADM systems. As a result, the proposal therefore pursues a **hybrid between a risk-based and a damage-based approach.**
- In return for this freedom, certain **due diligence and transparency obligations** should apply to the "high risk" category, in particular with regard to data quality and origin. ADM systems used by the public sector should meet even stricter requirements.
- **Effective supervisory and sanctioning mechanisms** are needed. In particular, Digital Society calls for state supervision of ADMS, deterrent penalties for companies and collective legal remedies for those affected.

Contents

1 Introduction

Artificial intelligence (AI) or automated decision-making systems (ADM systems, ADMS, see Glossary) are no longer futuristic wishful thinking. They are already in use in our everyday lives, helping us to make decisions, simplifying interaction with computers and generating complex texts, images or music. The positive potential is immense. For example, we will be able to talk to machines, tedious tasks will be further automated, and personalized medicine will revolutionize previous methods of diagnosis and medication. The US Food and Drug Administration (FDA), which is responsible for medical devices, already lists several hundred medical devices that use AI and machine learning (as of May 13, 2024, there are 882).[1]

But as with most technological revolutions, there are also negative aspects, and in the case of AI and ADM systems, they are not in short supply. Probably the best-known case is the child benefit scandal in the Netherlands, in which child benefit was wrongly reclaimed from thousands of families because, for example, multiple nationality was seen as an indicator of fraud.[2] Today, it is no longer possible to conclusively assess the areas in which artificial intelligence is used and those in which it is not. And we are only at the beginning of a change whose extent we cannot estimate. It is important to manage this development in such a way that we can benefit from the positive side while minimizing the negative effects.

Other states and associations of states – including the EU, China and the USA – have recognized the transformative potential of AI and ADM systems. They are actively trying to steer such systems in legally regulated directions. Digitale Gesellschaft has actively participated in these debates[3] and is also calling for a corresponding adaptation of existing regulations in Switzerland to the shifting balances and new challenges posed by AI and ADM systems, so that benefits and risks are well balanced.

In this document, we present our proposal for a legal framework for Switzerland. The proposal is technology-neutral[4] and follows a “human-centered”

[1] See Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices | FDA

[2] See the article: Aufsicht und Transparenz: Wie die Niederlande aus KI-Skandalen lernen (netzpolitik.org)

[3] For example, in the Council of Europe's framework convention on AI, human rights, democracy and the rule of law; see the dossier on ADM systems by Digitale Gesellschaft for more information

[4] Our regulatory proposal focuses on the effects and risks of ADM systems and not on banning specific technologies

approach: the systems should benefit people, i.e., people should be better off as a result of the use of AI and ADM systems.

The use of ADM systems must be linked to the conditions *of transparency* and *traceability*. An ADM system fulfills the condition of *transparency* if the bases of automated decision-making are disclosed. For example, has an ADM system that is used to assess the reintegration of people who have had an accident into the labor market recorded and evaluated all the relevant facts of the specific case? The criterion of *traceability* is intended to ensure that the decision-making of the ADM system is disclosed so that at least the individuals concerned are able to understand the assessment of the relevant facts – and challenge them if necessary.

The legal framework is based on an assessment of the risks posed by ADM systems. Affected individuals, authorized NGOs and a state ADMS supervisory authority should have the right to inspect applications that use ADM systems. The legal framework also takes into account the fact that the risk posed by a system can change over time. Furthermore, it is compatible with the new data protection concept of the digital society.

As a civil society, we no longer have the option of deciding whether or not we want ADM systems to be used. They are already a reality. However, we do have the choice of deciding in which areas we want to be supported by ADM systems and in which areas we reject such support.

2 Scope

Our proposal covers ADM systems that make decisions fully automatically with the help of technical systems or at least support them. We avoid the controversial term artificial intelligence and adopt the following definition of ADM systems from a recommendation by the AI Now Institute (Richardson et al. 2019, p. 20) to the City of New York:

> 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,[5] and the set of processes involved in implementing those tools.

In this way, we define ADM systems not from the perspective of the technology used, but from the perspective of the effects. This allows us to avoid the definitional problems that, for example, the OECD had to contend with (OECD 2024).

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 every AI or big data system should fall within the scope of this regulatory proposal. Furthermore, the statement that almost every computer program constantly makes decisions is true in principle, but not helpful from a regulatory perspective. The decisions covered by the regulation must be discernible as individual, discrete decisions and be of a certain significance. The decisions subject to regulation must therefore have an effect on the freedoms, life or health, economic or social situation of individuals or groups as individual, discrete decisions. This definition should also include nudging by an ADM system if its decisions cumulatively have a social impact.

If a technical system does not fall within the scope of the legal framework, no risk categorization according to section 6 is necessary, and the associated effort does not have to be expended.

3 Summary of the legal framework

The legal framework is a combination of a harm-based and a risk-based approach. In the first approach, sanctions are only imposed retrospectively in the event of harm, while in the second approach, high-risk applications are subject to certain conditions from the outset. Anyone using an ADM system must assess and categorize the risk it poses to individuals

[5] 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, behavior, and personal autonomy.

and society in the context of the protection goals. The legal framework provides three categories for this purpose: “low risk”, “high risk” and “unacceptable risk”, which are defined by the legislature. In principle, the categories are based on the risk posed by the system to individuals and to society as a whole. For example, “systems with low or no risk” pose little or no risk to individuals and none to society, while “systems with unacceptable risk” pose an unacceptably high risk to individuals or society. In between, there are “systems with high risk”. For these systems, a broad duty of transparency and due diligence applies, which should enable the public to assess the risk and thus their benefits. In contrast to systems with an unacceptable risk, those with a high risk are not prohibited.

First and foremost, we consider the impact of ADM systems on individuals. However, if easily scalable systems are used on a broad scale, a risk for society[6] can also arise that cannot be sufficiently measured or sanctioned by looking only at individuals. ADM systems that distribute individualized political advertising on social networks on a massive scale are an example of such a societal risk. In individual cases, the risk may be negligible with reference to individual sovereignty. However, on average, over many people, such ADM systems can have noticeable effects, for example on election results, and thus damage democracy. The Swiss case law, which has so far focused on individuals, such as the Data Protection Act, falls short in such cases. Our proposal addresses this shortcoming by recognizing these societal risks and proposes a remedy in the form of collective legal remedies such as class actions and a right of action for associations.

The legal framework distinguishes between ADM systems used in the private sector and those used in the performance of public mandates (see sections 7.1 and 7.2). For both, we call for a right of appeal for affected individuals, public ADMS supervision and authorized NGOs to guarantee the correct risk classification according to section 6 and the enforcement of the associated obligations.

The new ADMS supervisory authority (more on this in section 5.1) should collect complaints, check on the use of ADM systems in companies and public authorities on suspicion of their use, and be able to impose revenue-dependent administrative sanctions in the first instance. As a diverse expert body, composed of people with social science, technical and legal expertise, it should act independently and free from instructions, both financially and in terms of personnel.

In order not to hinder innovation, we rely on self-declaration instead of burdening companies and public administration with bureaucratic testing processes. This allows those affected to design the specific implementation of compliance with the rules within the parameters defined by the legal framework themselves. However, this freedom should be counterbalanced by additional obligations. These include transparency requirements, due diligence and the reversal of the burden of proof, as well as effective sanctions in the event of non-compliance.

The exact functioning of an ADM system is usually subject to trade secrecy. It is therefore difficult for outsiders to obtain evidence of the risk posed by a system. Therefore, if the accusation is justified, i.e. if a court enters into the action, there should be a reversal of the burden of proof.[7] In this case, the accused company must prove the correct classification as the first instance in the recourse chain. For systems in fulfillment of a public contract, we demand extensive transparency and publication of the systems and data (see Glossary), in line with the demand ”Public Money? Public Code!”[8]

The state ADMS supervisory authority supports the controller[9] in assessing the risks of ADM systems by providing checklists and good practice guides to ensure awareness and adequate handling. Anyone who misjudges the system they are using and thus fails to meet their obligations or operates a prohibited ADM system that carries an unacceptable risk should face severe and revenue-based penalties. These should be administrative sanctions that explicitly do not aim to punish individual employees through criminal law, since it is usually not an individual, but an organizational fault. However, these sanctions should remain a last resort.

[6] https://booksummaryclub.com/weapons-of-math-destruction-book-summary/

[7] The EU has recently introduced the reversal of the burden of proof for software products. For technically complex products, consumers can demand that a company disclose the necessary and proportionate evidence.

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

[9] According to the definition of Art. 5 lit. j DSG

An increasing number of internationally significant institutions, such as the European Union or the Association for Computing Machinery (ACM), are now addressing the need for regulation of ADM systems (see appendix, section A). We are convinced that the proposed legal framework will help to close the existing regulatory gaps. In the following, the core aspects of the legal framework – in particular the risk categories and the transparency and due diligence obligations – are explained in detail.

4 Societal relevance

A central insight is that **automated decision-making systems are neither objective nor neutral**, because they always represent the values of their developers and society and can therefore be regarded as a socio-economic mirror of a particular society. This is problematic because cultural and social values differ around the globe, whereas technologies such as ADM systems can be used across borders or globally. In addition, values can change over time, but the systems that were once defined and used in the long term do not necessarily have to implement this. Furthermore, ADM systems are tied to their function. Design decisions made consciously or unconsciously by developers have an effect on the way the system works and can have negative consequences. The function of a system determines a narrow scope of action that is often not questioned.

One example of this would be the use of ADM systems to reduce personnel costs in the social welfare sector. These systems make hard decisions about the available resources of people dependent on social welfare. On the one hand, the basis for these systems' decisions is questionable, because the goals developers strive for and the underlying social values can change over time. On the other hand, the use of such systems is fundamentally questionable: recent scientific studies suggest that an increase in personnel costs tends to reduce social costs overall[10] (Eser Davolio 2020), which could argue against the automation of the corresponding administration.

However, people are often unaware of these limitations. Instead, the results of ADM systems are all too often considered to be objectively correct and accurate. In the context of decision problems, however, this view is deceptive because **there is no optimal solution for many decision problems**. However, by delegating everyday tasks to an ADM system, interactions with it and its results become part of social reality. Sociologist Michele Willson describes this as follows: “An algorithm is given a task or process, and the way it is used and handled in turn affects the things, people, and processes with which it interacts – with varying consequences” (Willson 2017: 139). Feedback effects arise that cause automated decision-making systems and their (sometimes erroneous or inaccurate) data bases to constantly change and become part of the social fabric. Such effects can be both intended and unintended.

One particularly problematic effect is discrimination. ADM systems trained on data sets reproduce the implicit discriminatory practices contained in them, for example against socially weaker groups or people with disabilities. For example, recruitment systems trained on historical data would probably continue to discriminate against women or people with disabilities more often, even though this is no longer tolerated. Humans do not necessarily make better decisions and are not free of prejudice. However, they can reflect on these, also with the help of technology, and exchange ideas. These systems **lack this ability to reflect,** which is why they should not be delegated decisions that have lasting effects on society. The discriminatory effects of ADM systems can be exacerbated, especially by their increased use across borders.

Another problematic effect is that automated decision-making systems are increasingly curating the information overload, for example in the form of so-called “suggestion systems” or as “fact” and “copyright” checkers. This empowers system operators to selectively reinforce political messages and positions through targeted agenda setting. These systems (e.g. fact checkers) do not necessarily have to operate on personal data. Many of these effects of automated decision-making systems have in common the fact that they often occur covertly and their effects are only noticed late or through further, indirect effects. The use and functioning of the systems are often not known because their developers and operators have no interest in disclosure.

Finally, the interaction of different ADM systems gives rise to additional risks that are difficult to assess. The emergence of feedback-reinforcing ef-

[10] See https://www.zhaw.ch/de/forschung/forschungsdatenbank/projektdetail/projektid/1668/

fects (feedback loops, see Glossary) is foreseeable and thus a societal risk. However, this **increase in complexity** is still faced by people who have increasing difficulties in penetrating it, even with full transparency of the individual systems.

In order to be able to assess these effects to some extent, a far-reaching **transparency and due diligence obligation** should therefore apply. At least in the case of important automated decision-making systems, a space should be created for a sustainable, public discourse on the norms and values that underlie the metrics, measurements or key figures that are created, evaluated and interpreted. **People should have sovereignty over ADM systems,** and not the other way around.

Furthermore, many effects are not clearly comprehensible from an individual perspective and only become visible through the accumulated observations of many affected parties. Unfortunately, however, the existing laws usually argue from a single-case perspective. Therefore, we need methods for **collective enforcement**, which have rarely been found in Swiss law so far.

By focusing on the effects and risks, a **technology-neutral formulation allows** us to react flexibly to new methods or changed possible uses of existing technologies. The goal should be that people fare better through the use of this system. These and other topics are being discussed intensively in the scientific community, but in this section of the position paper they are only outlined for the sake of completeness.

5 A regulatory proposal for ADM systems

As a society, we are increasingly confronted with the specific effects of ADM systems. Therefore, we demand **a substantial expansion of existing laws** to take into account the challenges of automated decision-making systems or even a specific **"ADMS act"** if this proves advantageous. We demand **transparency** in the use of ADM systems in order to be able to apply the already existing laws and **effective penalties** in case of disregard. We demand **explainability** of ADM systems, which is quickly and with appropriate cognitive effort accessible to humans. We demand a constitutional control of critical ADM systems with the possibility to intervene if necessary.

We see concrete effects primarily on individuals and want to give them opportunities to enforce their rights. However, there are risks that tend to affect society as a whole, such as political influence through personalized advertising campaigns or self-reinforcing feedback effects, in which linked systems – with or without human intervention – could form their own value cycles. **Transparency is key, but not sufficient without further measures.** The right to informational self-determination requires not only knowledge of the processes, but also the possibility to exercise some control over them.

We demand **clear protection goals, namely the observance of fundamental and human rights, the protection of the mental and physical health and safety of the individual, the protection of life and development opportunities, as well as the protection of democratic rights and processes**. Furthermore, the persons concerned and the public must have the opportunity to effectively monitor compliance with these protection goals and, if necessary, to object to them and demand them at a low threshold. The human being should have the sovereignty of validity, he should therefore generally stand above the machine with his interpretation and be able to achieve his ideas and goals better, faster and with fewer errors through ADM systems.

ADMS regulation should neither inhibit innovation nor place a disproportionate bureaucratic burden on companies and the ADMS supervisory authority, which is described in detail later. We advocate a broad legal framework that generally introduces the necessary regulation but allows 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 below, is **compatible with the European Union AI Act**, but, unlike the EU's application-based categorization, allows for a context-dependent

[11] We demand an adaptation of Art. 21 para. 1 DSG (deletion of "exclusively"): "The person responsible informs the person concerned about a decision that is based on automated processing and that has a legal consequence for them or significantly affects them (automated individual decision)."

[12] In terms of fairness in relation to decision algorithms, it is about the evaluation and correction of algorithmic bias. Outputs of decision algorithms are considered "fair" if they are independent of specific variables such as gender, age, etc. However, the exact (mathematical) formulation of fairness is still an open debate, and some definitions even contradict each other.

classification of applications. An overview of other regulatory efforts, both in Switzerland and internationally, can be found in Appendix section A.

To support the responsible development of ADM systems, government measures should be taken to promote open source libraries and frameworks for ADMS and AI developers.

In principle, existing laws can also be applied to ADM systems with some modifications. For example, the dispersion and use of personal data can be regulated by the Data Protection Act.[11] Prohibitions on discrimination are the negative of the emerging fairness discussion,[12] however, there is a large gray area in between[13], where the majority of real-world applications will be found. Labor and data protection laws prohibit some monitoring practices, algorithmic and non-algorithmic, in the workplace.[14] From a formal legal point of view, a separate ADMS act could prove advantageous for two reasons: firstly, because the necessary amendment in other legal texts requires a common definition of terms, categorization and risks, and secondly because the ADMS supervision that follows can hardly be defined elsewhere.

Legal entities can benefit from ADM systems in the same way as natural persons. The use of AI, for example, in production, service provision and so on is manifold. At the time of publication, Digital Society is not aware of any specific examples that indicate discrimination against legal entities. Nevertheless, Digital Society is aware that legal entities may also experience disadvantages. Whether and to what extent this can happen and to what extent existing laws (such as competition law) insufficiently protect legal entities cannot be estimated at this point in time. Digital Society reserves the right to comment on this issue at a later date.

5.1 The ADMS supervision

The **state-run ADMS supervisory authority** is to act **as a centre of expertise**. It advises companies, authorities and the public and orchestrates any long-term analyses. It collects complaints from affected parties and, independently of instructions, monitors compliance with the regulation and the categorization of state and private-sector ADM systems if there is sufficient suspicion.

The ADMS supervisory authority is to be authoritative at all levels (federal, cantonal and municipal). It can impose first-instance sanctions in parallel with the complaints and legal channels of individuals and authorized NGOs in the event of violations. The authority to monitor compliance with ADMS regulation and to prevent and sanction risks to individuals and society is concentrated in this agency, with uniform public responsibilities for the entire public sector at all levels (see section 8).

It supports the assessment of the risks of ADM systems by providing checklists and good practice guides to ensure awareness and adequate handling of the issues. It should be a diverse authority (composed of individuals with different social science, technical and legal expertise) that acts independently of instructions and with its own budget, for example like the Federal Data Protection and Information Commissioner (FDPIC). How its supervisory activities should be implemented in practice in line with the principles set out here to ensure the best possible protection for individuals and society remains to be determined in detail.

5.2 IT security

Automated systems, like other IT systems, are never 100% secure and may be vulnerable to hacking or misuse. Their functions can thus potentially be manipulated by insiders or third parties. However, we believe that measures to prevent such attacks do not belong in an ADMS regulation, but in a general "**IT security law**".

The ADMS law should instead deal with the specific effects of automated decision-making systems. In doing so, risk classification must not only take into account the intended use of the system, but

[13] While discrimination as an offense presupposes a serious violation of fairness, perfect fairness is usually achievable only for a specific metric, requiring the neglect of other equally valid metrics. In between, there exists a vast gray area.

[14] The monitoring of employees is allowed to a limited extent, such as the recording and adherence to working hours (Art. 46 ArG), data related to suitability for the employment relationship et cetera. The limits of surveillance are found in the protection of privacy (Art. 328 et seq. OR), data protection according to DSG and in certain mandatory articles of the labor law. Systematic monitoring of employee behavior is not permitted (Art. 26 para. 1 ArGV 3), as it can have health effects on employees. Exceptions may be permitted (Art. 26 para. 2 ArGV 3) if they are made for other reasons, such as optimizing performance or quality assurance, and only if proportionality is maintained and the risk to personality and health is minimized (case-by-case assessment) (cf. Bürgi and Nägeli 2022).

[15] This includes, on the one hand, the unlawful use or "hacking" of these systems, but also ADM system-specific effects, such as extracting sensitive training data from the models (see glossary) themselves, the unreliability of predictions for data series that have not been used for testing (fragility), specially generated data series that look correct to humans but result in incorrect outputs (adversarial examples), etc.

must also consider foreseeable, possible incorrect and abusive uses[15] ("reasonably foreseeable misuse", EU AI Act Art 9.2(b)). One example is the analysis of communication between all employees for the purpose of improving collaboration. A foreseeable misuse of this system is the monitoring and/or evaluation of employees.

An important building block for the reliability of algorithms is the application of **product liability to software** and thus to all types of computer algorithms. It should not be possible for software companies to evade responsibility by cleverly formulating terms and conditions. Due to its generality, however, product liability should be part of such an IT security law and not only defined specifically for ADM systems.

6 Categorization

Our proposal follows a mixed form between a damage-based and a risk-based approach. In the case of a damage-based approach, sanctions are only imposed retrospectively in the event of damage. In risk-based regulation, applications are subject to appropriate conditions from the outset. In doing so, we follow the recommendations of the German government's "Report of the Data Ethics Commission" (page 43ff)[16] and the detailed analysis of the fundamental rights implications of facial recognition technology in FRA 2019). As a result of this hybrid form, applications that are high-risk and high-impact are subject to due diligence and transparency requirements from the outset, while we rely on self-declaration for applications that are lower-risk and lower-impact. Potential breaches of duty or miscategorizations are penalized a posteriori through penalties in the context of complaints and lawsuits. This approach gives the operators of ADM systems the opportunity to develop and implement ADM systems independently, but within a clear framework. The penalty mechanisms demand and strengthen personal responsibility.

We divide ADM systems into three categories: "low risk", "high risk" and "unacceptable risk". Automated decision-making systems are categorized according to the risk they pose to individuals – the individual case perspective – and to society. The risk to society is determined in the context of the protection goals based on the potential for harm and the probability of occurrence for society as a whole, while the risk to individuals is considered in each individual case. The decision tree for the categories is determined by the legislator, not by other actors.

We discuss these risks in more detail in the next section. We then explain the assessment criteria according to which ADM systems should be categorized. Finally, we discuss the specific categories.

6.1 Protection goals and risks for individuals and society as a whole

When ADM systems are used, the protection of individuals and society must be ensured. The fundamental protection goals with regard to individuals are compliance with fundamental and human rights. In this regard, we are guided by the protection goals of our data protection concept (cf. Digitale Gesellschaft 2023), which are primarily designed to protect individuals but also include societal risks that are not directly based on the accumulation of individual risks:

- Protection against manipulation
- Protection against discrimination
- Protection against surveillance and the right to anonymity
- Protection against adverse effects on health, life and development opportunities
- Right to transparency and duty of care
- Right to be forgotten
- Protection of the open society and free democracy

Manipulation is to be understood as the intentional, targeted and usually covert influencing of another person's decision in order to undermine their self-control and decision-making power. ADM systems allow a highly automated and individualized approach to individuals. Manipulation can lead to a disadvantage for the person concerned. It aims to control the behavior of individuals or groups by exploiting human weaknesses. Vulnerable people are particularly at risk.

Discrimination occurs when ADM systems disadvantage people or groups based on characteristics such as ethnicity, skin color, gender, class, sexual orientation, etc. In most cases, the underlying "bias"

[16] See the German Federal Government's Data Ethics Commission 2019

is already present in the training data and is perpetuated or even amplified by the automation of decisions. Examples of discrimination are given in the dossier of the Tracking & Profiling expert group of Digitale Gesellschaft.

If the protection against surveillance and the right to anonymity are violated, people are prevented from developing their identity. There may also be "chilling effects", i.e. the mere possibility of surveillance may cause people to refrain from participating in demonstrations and rallies, with extremely negative effects on the democratic decision-making process of a society (this is the main reason why net policy NGOs are vigorously fighting against biometric identification and facial recognition in public places).

ADM systems that make or support decisions in the areas of social services, law enforcement, education, and daily economic life can have a massive impact on individuals' development opportunities (for example, not being granted a university place or a loan). For examples and further details, please refer to the dossier of the Tracking & Profiling section of Digitale Gesellschaft.

The right to transparency can be violated at several levels. For example, individuals may not be aware that decisions concerning them are being made in an automated manner. For example, according to the new Data Protection Act, only fully automated decisions must be reported. Furthermore, it is generally not transparent to individuals which data is used or which models and coefficients (see Glossary) are used to make decisions. On this point, we also refer you to the dossier of the Tracking & Profiling specialist group of Digitale Gesellschaft.

The right to be forgotten is rarely discussed in connection with ADM systems, but it is also relevant there. How long may data from the past be considered for automated decisions, such as when calculating credit scores? This is particularly challenging if an ADM system has already been trained on this data and would have to be corrected.

Open society and free democracy are threatened in the context of ADM systems where people or entire groups are discriminated against on the basis of certain characteristics (see protection goal of discrimination) and where democratic processes are disrupted (see protection goal of protection against surveillance and right to anonymity). Democracy is also threatened when people or entire groups are manipulated – for example, in social media. Or when messages are targeted and individually tailored to specific groups of people, and there is a lack of transparency regarding what information is being displayed. This fragments the discourse space and makes it more difficult to hold a substantive debate.

6.2 Assessment criteria

Risks are understood as a combination of the severity of the possible damage and the probability that the damage will occur. In short, risk = extent of damage * probability of damage occurring. The damage is assessed in terms of the protection goals (see 6.1). The probability of damage occurring can result from several factors. One way of evaluating this would be to consider, for example, how likely it is that a problematic situation will arise (exposure) and how likely it is that this can be corrected before the damage occurs. In some circumstances, the reversal of the damage must also be taken into account, for example, a subsequent money transfer after an initial block.

For the settlement, ADM systems should be categorized according to their risks. For the categorization, we adopt and supplement some of the concepts from the EU Commission's AI Act (Art 7.2). When assigning an ADM system to a risk category, the following aspects should be considered:

- What are the purpose and scope of the ADM system?
- To what extent will the ADM system be used (selectively or across the board)?
- To what extent is there known harm to health, harm to safety, or adverse impacts on fundamental rights that have resulted from the use of the ADM system? Is there significant concern about the occurrence of such harm, such adverse impacts, or such adverse effects based on reports or documented allegations that should be communicated to the appropriate authorities?
- What is the potential magnitude of such harm, damage or adverse effect, particularly in terms of its intensity and its potential to impact a wide range of individuals?
- To what extent do persons who are potentially harmed or adversely affected depend on the output produced by an ADM system and to what extent do they rely on the ADM system because, in particular, it is reasonably impracticable or legally impossible to avoid using the ADM system?
- To what extent are potentially harmed or adversely affected persons vulnerable vis-à-vis the entity deploying an ADM system, in particular due to an imbalance in terms of power,

knowledge, economic or social circumstances or age?

- To what degree and how easily can the result produced by an ADM system be reversed? Results that affect the health or safety of individuals cannot be considered easily reversible.
- Can destructive or self-reinforcing feedback loops arise, and what measures are taken against them?

If the application of an ADM system to an individual can be avoided by that individual, assuming average knowledge and normal circumstances and without incurring any disadvantages, then that system falls into a lower category than if an individual is dependent on that system. If the effect of an automated decision can be (easily) reversed (or compensated), then this system also falls into a lower category. The prerequisite for this is that individuals are not only aware of the automated decision itself, but also of the possibility of appeal and reversal, and that this reversal can be requested with normally assumed knowledge and without incurring disadvantages in a timely manner.

6.3 Categories

If a specific technical system falls within the scope of this Act (i.e. it is an ADM system as defined in section 2), it should be classified in one of the following three categories: **"low risk**", "**high risk**" and "**unacceptable risk**". The due diligence and transparency obligations set out below apply only to 'high risk' systems.

In this classification, the risks are considered in terms of the protection objectives in accordance with section 6.1 and the assessment criteria in accordance with section 6.2 are applied. The assessment of whether an ADM system is involved and what risk is associated with it is carried out by the entity using the ADM system on a self-declaratory basis. This is to keep the administrative burden as low as possible. In the event of an incorrect or insufficient self-declaration, there is a risk of high and turnover-dependent administrative sanctions, depending on the degree of culpability. The assessment will therefore ultimately fall to the courts.

At the same time, in order not to inhibit innovation, it is also important that legal certainty is maintained in the self-declaration. The administrative sanctions should not affect companies that carry out the self-declaration carefully and to the best of their knowledge and belief. This circumstance should be taken into account in particular by the ADMS supervisory authority in the form of information sheets that specify the criteria for self-declaration and provide examples (see, for example, in data protection law: FDPIC 2023).

This risk-based categorization is compatible with the application-based formulation of the European Union's AI Act (EU AI Act). However, in contrast to the EU AI Act, we do not fundamentally prohibit applications, but consider them in the light of the respective circumstances. For example, algorithmic emotion recognition may be prohibited in job interviews, but an art exhibition may use it because the risk to society and individuals is low in the second case. The risk of automated decision-making systems and thus their assessment can also change over time and with the development of technology and society, as well as in interaction with other systems. The proposed categorization scheme can reflect these developments.

The lowest category ("low risk") includes systems that

- pose a low risk to society, and
- for individuals
  - pose no (or only a minor) risk to the protection goals.

The systems in this category are therefore characterized by the fact that they are unlikely to have any particular negative impact on individuals or society. If medium-level damage or encroachment on fundamental rights is possible, but the system can be easily avoided without the need for extensive specialist knowledge, or if harmful effects can be easily reversed, a system can still be classified in this category. Systems whose decisions can have a damaging effect on the health and/or safety of individuals cannot, in principle, be placed in this category.

Examples of ADM systems in this category are the automatic inspection of food packaging for correctness directly after production or ADM systems for predicting pollen levels, which are of great help to people with allergies but have only a negligible impact in the event of malfunction.

The middle category ("high risk") includes systems that

- represent a high risk for the company; or
- for individuals
  - pose a high risk to the protection goals.

The systems in this category are characterized by the fact that their (positive) benefits are offset by a significant potential for negative outcomes. The potential for harm is still acceptable (or can be mitigated), otherwise such a system would be classified in the

next higher category. Systems in this category are typically used widely (not just occasionally) and do not allow individuals to opt out of automated decision-making. In this category, damage caused by decisions cannot realistically be reversed or compensated for either.

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

Individual decisions in the social services (e.g. eligibility assessments) are highly risky for several reasons: they usually affect vulnerable people, cannot be avoided/circumvented, and it is generally not possible for those affected to obtain corrections of wrong decisions without complicated and expensive legal action. Decisions that lead to the recruitment or selection of individuals in job application processes also have a high risk because they cannot be avoided by those affected, but they influence the life and development opportunities of these people.

In addition, there are systems with the potential to have irreversible and serious effects on individuals, for example in medical diagnostics. These would thus be categorized as "unacceptable". As long as these systems are used as support systems and the final decision is made by a professional, a downgrade to "high risk" is reasonable. However, there is a fluid transition from recommender systems that are frequently used as support systems to unquestioned acceptance of these suggestions, which could cause initial support systems to mutate into de facto decision-makers.

The highest category ("unacceptable risk") includes systems that

- pose an unacceptable risk to society as a whole; or
- for individuals
  - an unacceptable risk to the protection goals.
  - represent irreversible and serious effects.

This category includes systems whose potential damage is so great that it cannot be risked. For many systems in this category, the damage is also known and documented, and thus no longer potential, but can be reliably expected. Furthermore, the decisions in this category are neither reversible (e.g. biometric mass surveillance) nor revisable, and often cannot be verified either (e.g. automated/supporting asylum, probation or court decisions). The expected or proven harm to individuals and society is so great in this category that the risks cannot be accepted or mitigated. The use of such systems is prohibited.

Examples of unacceptable effects for society are the aforementioned biometric mass surveillance (including facial recognition[17]), which not only represents a massive encroachment in fundamental rights such as human dignity, autonomy and privacy, but also has a chilling effect on democratic processes and thus on society (c.f. Assion 2014 and Penney 2016). Another example is the automated evaluation of behavior (social scoring), which primarily affects individuals but can have far-reaching (and, in addition, not democratically legitimized) effects on society due to its control and shaping effects.

For individuals, we see unacceptable effects not only in asylum, probation or court decisions, but also in the surveillance of employees, students and pupils. Far-reaching automated assessment in the workplace and the resulting dismissal or optimization decisions can cause unacceptable harm to the physical health of employees.[18]

7 Due diligence and transparency obligations

In principle, the entity that uses the ADM system from the perspective of the data subject (e.g. the operating company) is responsible for its functionality and its correct classification in the above-mentioned risk categories. While it should be possible to pass on certain business risks to the manufacturers of com-

[17] For example, https://gesichtserkennung-stoppen.ch, https://reclaimyourface.eu

[18] For these reasons, there have already been calls to add surveillance and automated management in work and educational contexts to the list of prohibited applications (cf. EDRi 2021); see Crawford et al. on the criticism of techniques of automated emotion recognition (cf. Crawford 2021)

[19] This implies a clearly verifiable functionality of the autonomous decision-making system.

ponents or systems under civil law, it should be prevented (via the product liability of the separate IT security law mentioned above) that manufacturers can release themselves from all responsibility by means of the general terms and conditions, as is currently common practice in software usage contracts.

We consider a certification requirement to be useful only in special and private-sector areas of application with a specific and easily standardized purpose[19], such as for medical products, for example an automatic defibrillator (AED). Otherwise, there is a risk that accountability will be outsourced to certificate issuers on a large scale.

One way to support due diligence would be impact assessments, which shed light on the possible consequences of the development and use of an ADM system and lay the foundation for well-thought-out risk mitigation strategies, as well as serving as an indication of the responsible use of the technology. However, at the time of writing, the Digital Society does not see any direct advantage in making these instruments mandatory.

The classification into a risk category is to be documented by the operator. For ADM systems with "low risk", an informal but comprehensibly justified classification is sufficient. For ADM systems with "high risk", a systematic analysis is necessary, which can be carried out, for example, as part of a risk management process that must exist for certain product classes such as medical devices anyway.

The following transparency obligations apply only to the ADM systems in the "high risk" category. "Unacceptable" systems must not be used from the outset. A false or insufficient declaration is punishable by severe penalties. The degree of transparency obligations should enable the assessment of the individual systems with regard to their risks, but also provide sufficient information to assess the effectiveness of the entire ADMS ecosystem. We therefore recommend standardized transparency reporting formats.

We distinguish between obligations for systems used in the **private sector** and those used **in the performance of a public mandate**. In general, a **labeling and notification requirement** applies to all systems (private and public) categorized as "high risk"

1. indicates that an ADM system is being used,[20]
2. a short abstract on the purpose of the system and specific possible outputs, as well as
3. information about the origin of the data, as well as explanations of the specific features (see Glossary) used by the ADM system and what these represent.

The information on the origin of the data should also serve to ensure that the data quality is sufficiently high and that the **chaining** of several ADM systems (possibly from different manufacturers) is more visible. Furthermore, we call for a periodic and continuous review of the risk (i.e. the classification of the category) and the documentation regarding transparency obligations, especially for self-learning systems.[21]

The **data quality** mentioned in the previous paragraph is about ensuring that the data used either matches reality[22] so that systems based on them function as flawlessly as possible, or that they have only been modified in intended and generally useful aspects, for example to prevent discrimination.

With regard to the **source of the data**, the right to informational self-determination must be respected; this also applies to data from abroad. Data must come from ethically justifiable sources; for example, the use of data obtained illegally is to be avoided as a matter of principle.[23]

Furthermore, it should be clearly stated when data from other ADM systems is used. This is to create transparency regarding the **interlinking of** such systems, which create their own risks due to the foreseeable increase in the complexity of their interaction, their (opaque) information flow and the resulting feedback loops.

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

[20] With an amendment to Art. 21 para. 1 nDSG (deletion of "exclusively")

[21] These are systems that continuously adapt their functionality based on new input. This functionality leads to a variety of problems, such as systems unlearning previously attested guarantees such as "fairness", or being fed manipulated data intentionally and in a way that is difficult to detect, in order to change their functionality to their own advantage.

[22] For example, it is statistically representative (with regard to the system's purpose and area of application), accurate, complete and as consistent as possible and follow a known semantics

[23] Or rather, its use is only justified under certain circumstances from an ethical point of view after weighing up the pros and cons (cf. Imhasly 2021).

7.1 Private sector context

For entities that use systems in a private-sector context, we require information about the origin of all data used and about the quality and completeness with regard to the purpose of the ADM system. This includes all data used for setting up, training, validating and predicting the system, etc. It also includes documentation on the purpose of the system and meaningful information about which features are used as input to assess the scope for individuals and society, and the risk to the protection goals, in particular the health, safety or fundamental rights, of the individual or society.

In weighing risk and benefit, formal regulation can provide for exceptions to the transparency obligations as well as liability for standardizable products if a certification body is created at the same time that fulfills the above-mentioned quality requirements at least equally well (for example, in medical diagnostics).

7.2 In fulfillment of a public mandate

For state actors, a higher standard for traceability and transparency already arises from existing regulations, for example from the right of publicity and the code of criminal procedure. These standards also apply as a minimum when using ADM systems. In addition to the same transparency obligations as for private-sector ADM systems (information on data origin and quality, features and purpose), we require the disclosure of coefficients (see Glossary) in a standardized format[24] for ADM systems of state actors.

- For ADM systems based on non-personal data, the data should be made available as open data as far as possible, together with the coefficients.
- For ADM systems based on personal data or 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 coefficients; or b) neither the data or the coefficients are published if (in exceptional cases) the generation of synthesized data and the use of corresponding ADM systems is associated with disproportionate effort or if personal data or non-personal data that may not be published can be derived from their coefficients. However, in this case, the ADMS supervisory authority and authorized NGOs must be given access to review the implications for individuals and for society and the risk to the health, safety or fundamental rights of individuals or society.

The general disclosure requirement called for here corresponds to the “Public Money? Public Code!” requirement – the requirement for source code disclosure of software financed by public funds. This also allows the solutions to be used and further developed by other authorities or the public.

The guidelines of the federal government for artificial intelligence and their monitoring show the relevance of the topic for the federal administration.

8 Controls, measures, and sanctions

Violations of the due diligence and transparency obligations listed above should be effectively sanctioned. Here, too, we distinguish between private and public use. In both cases, compliance with the due diligence and transparency obligations is monitored by individuals on the one hand and by authorized associations (NGOs) on the other, which can file complaints or lawsuits in the event of damage. Associations should be entitled to file complaints if they are active throughout Switzerland and have a corresponding purpose enshrined in their statutes. The possibilities for control, measures and sanctions, as well as the avenues for legal redress, 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 supplemented or redesigned. This also includes the revision and improvement of collective redress mechanisms.[25]

The ADMS oversight body should be able to investigate violations of the regulation ex officio and formally issue orders. It can demand access and impose sanctions. To ensure the best possible protec-

[24] This simplifies automatic inspection.

[25] At the time of publication of this document, the introduction of general collective redress mechanisms in the Swiss Code of Civil Procedure (ZPO) is being debated in parliament. While an anchoring in the ZPO would be welcome, we call for the introduction of collective redress mechanisms regardless of the outcome of these deliberations.

tion for those affected, both intentional and negligent actions should be punishable. We expect that questionable systems will quickly come to light, drawing the attention of civil society and thus of the relevant associations or the ADMS oversight body.

Misclassification of ADM systems, for example as low risk instead of the correct high risk, and the associated violations of due diligence and transparency obligations, should be prevented by imposing sufficiently high sanctions. For this reason, we consider the proposed self-declaration requirement to avoid bureaucratic processes and to relieve the burden on companies to be sufficient. We expect companies to implement their own rules along the lines of data protection, for example by setting up internal reporting offices for suspected violations or false declarations.

8.1 Private sector

Proving individual guilt does not seem to be effective, since violations of the due diligence and transparency obligations listed above usually involve organizational fault. ADMS oversight should therefore punish the companies by means of administrative sanctions and not sanction individuals through criminal law. This also eliminates the otherwise looming "shifting" of blame onto "scapegoats". Furthermore, the range of penalties must be dependent on turnover so that large companies cannot get off comparatively cheaply. 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".

Underestimating the category of the ADM system is punishable.

The ADMS supervisory authority has the following instruments at its disposal: it collects complaints, it can demand access and impose sanctions and issue rulings. The courts have the final say.

In the event of suspected inadequacy, we see the following avenues for redress:

1. Affected individuals can file lawsuits against private sector entities and complaints against orders issued by the ADMS supervisory authority if the orders issued by the ADMS supervisory authority are considered inadequate. Class actions and class complaints should also be explicitly possible. The legal remedies are to be adapted in this sense.
2. Authorized associations (throughout Switzerland and with a suitable purpose according to the statutes) should be able to file a lawsuit against private entities and complaints against orders of the ADMS supervisory authority without being personally affected (right of association to file complaints or lawsuits). In view of the high costs of litigation and as a regulatory element, the relevant association can receive a portion of the sanction amount as compensation for expenses.

In the event of an impending conviction, the defendants' interest in concealment leads to a strong imbalance of power. In the event of a serious accusation, that is, if a court recognizes the complaint or action as admissible, we therefore demand a **reversal of the burden of proof** so that accused entities (operators of ADM systems) must sufficiently prove that they have not violated the categorization requirements, due diligence or transparency obligations. This reversal of the burden of proof is one of the reciprocal obligations for the advance of trust in ADM self-categorization by companies.

As with similar technology and software products, the operator should be able to claim compensation for damages attributable to its suppliers, such as developers or system operators. However, the operator itself always remains responsible for the data subjects.

8.2 In the performance of a public mandate

In principle, the same controls, measures and sanctions should be possible as against private entities. How the relationship between ADMS oversight and the entities in the performance of a public mandate is to be structured in detail at the cantonal and municipal level remains to be clarified.

There should be options for both individuals and associations (analogous to section 8.1 Private sector) to take action against risks posed by ADM systems and against the results of such systems. In order to avoid any conflicts of jurisdiction, for example, when entities with a public mandate at the municipal or cantonal level are affected, the ADMS supervision in the cantonal proceedings could always be granted the rights of a party.

Here too, as is usual with similar technology and software products, it should be possible for the authority as operator and client to claim compensation for damages attributable to its suppliers, such as developers or system operators. However, the authority always remains directly responsible to those affected.

9 Future considerations

In the future, systems for automated decision-making will take over more and more tasks, work and functions, which may result in new consequences, opportunities, challenges and problems. In the following, we would therefore like to address various points in the sense of a technology impact assessment for which regulatory intervention could be 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. It is therefore important to take a very close look at how these dependencies develop.

The second concerns the **networking and interlinking** of wide-ranging automated systems: in the near future, the output of one system could partly be the input of the other system, which in turn could have an influence on the first system. This can lead to complex and multi-layered feedback effects, especially with more than two systems, and thus to risks that are difficult to assess. The foreseeable partial lack of transparency 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 workings 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 the coupling of systems above a certain cluster size or if certain security or purpose criteria are no longer met.

In various discourse 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 world view attempts to squeeze the complete reality into a construct of formulas and numbers. Based on our insight that there is no such thing as absolute objectivity and that all metrics, measurements and key figures, as well as their interpretation, are therefore subject to social negotiation processes, we see this path as misleading. We therefore advise general **data minimisation as a basic principle**, since, as with data protection, it reduces the problems that arise at the source.

Furthermore, a **dependence on ADM systems** is foreseeable. The use of automation to facilitate and reduce the workload makes it possible to accomplish more and more complex tasks in less time. However, we should be aware of what a failure of these automated systems would mean for us, what range it would have and what risks would be involved, and prepare quickly implementable emergency strategies as a measure. Increasing networking and dependence on individual resources, as in the case of the internet, also increases the risk that many functions could fail simultaneously. Perhaps it makes sense to talk about measures that produce completely redundant systems.

At the same time, one can also foresee a potential **loss of competence in humans** and a **loss of accountability**. Delegating tasks to automated systems, relying on them to be carried out correctly and the associated habituation effects could lead to a loss of skills that are not needed without the corresponding systems, and to a lower level of individual or collective accountability. Such effects may only become apparent over the course of several generations, for example if certain skills are no longer passed on.

Simple jobs without long training requirements will increasingly be lost. This can have significant social consequences, including in Switzerland, where, among other things, social status and work are closely linked. We need to reflect on our understanding of social esteem and thus also on the distribution of wealth, and possibly redefine it in the long term so that **all people can share in the benefits of automation**.

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 federal government, among other organizations. The general goal of technology impact assessment is to systematically analyze and evaluate the effects and consequences of technologies in all visibly affected areas of the natural and social environment.

A Regulatory proposals for ADM systems, artificial intelligence, and algorithms

In the course of digitalization, (data-driven) information systems have spread virtually unregulated (aside from mainly European data protection laws). Depending on the aspect to be emphasized, these systems are referred to as “automated decision-making systems”, “algorithmic systems”, “artificial intelligence” or “big data systems”. Despite their differences, these systems have in common the fact that they process and analyze very large amounts of data 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 most significant negative effects and risks. 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) developed a position paper on the transparency and accountability of algorithms as early as 2017 (cf. ACM 2017) and updated it in 2022 (cf. ACM 2022). The ACM is the professional association of US computer scientists and thus the organization to which many of the people who are at the forefront of research, development, and deployment of ADM systems belong. Many of the statements and principles in this positioning, for example regarding transparency and data, are also reflected in our proposal.

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

Many previous approaches and proposals (cf. DEK 2019, EU AI Act 2021) take a risk-based approach, in which an attempt is made – as our proposal also does – to divide ADM systems into categories based on their intrinsic risk and to define stricter rules for the categories with increasing risk. The number of categories varies between the proposals. However, there is always a (risk-free or low-risk) lowest category with very few rules or requirements, and a category of ADM systems classified as very risky, the use of which is prohibited. The Data Ethics Commission's proposal, for example, developed the risk pyramid in this way.

The EU Commission's proposal “AI Act” (cf. EU AI Act 2024), adopted in June 2024, also follows this risk-based approach. In contrast to our proposal, the AI Act contains specific 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 discussion. While there is broad agreement on the necessity and relevance of this proposal and its general, risk-based approach, there is also detailed criticism from civil society (for example, Digitale Gesellschaft, with a large alliance of the EDRi network, is calling, among other things, for a broader version of the prohibited and high-risk categories and for the deletion of exceptions, particularly in the area of biometric identification; cf. EDRi 2021). Similar criticism is also coming from consumer protection organizations (cf. VZBV 2021).

The “AI Bill of Rights” (White House Office of Science and Technology Policy 2022) in the US is, strictly speaking, not a separate regulatory attempt, but rather formulates and sharpens fundamental rights in the context of artificial intelligence, such as the right to protection against algorithmic discrimination, the right to transparency and explanation, or the right to human intervention. Although the bill has “AI” in the title, many of its statements apply particularly to ADM systems. The Accountability Act is another proposed law from the US. This proposal covers critical decisions about consumers, for example in the areas of education, work, or healthcare. The proposal aims to minimize negative effects on consumers and proposes various rights for consumers, such as labeling requirements for ADM systems, opt-out options, and opportunities to object and correct.

In the PRC, too, there are proposals for dealing with AI and ADM systems (cf. National New Generation Artificial Intelligence Governance Specialist Committee). This proposal formulates ethical guidelines in the field of AI. The guidelines are grouped into basic standards such as promoting human well-being, promoting fairness and justice, protecting privacy, and strengthening accountability.

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

B Bibliography

B.1 Sources regarding regulatory proposals for ADM systems in a European and intercontinental context

- ACM (2017). Statement on Algorithmic Transparency and Accountability (accessed 13.0.7.2024). *ACM U.S. Public Policy Council.*
- ACM (2022). Statement on Principles for Responsible Algorithmic Systems (accessed 13.0.7.2024). *Europe/U.S. Public Policy Council.*
- BAKOM (2022). Monitoring der Leitlinien «Künstliche Intelligenz» für den Bund (accessed 13.0.7.2024). *Bundesamt für Kommunikation BAKOM*.
- Bundesrat (2020). Die Leitlinien des Bundes für Künstliche Intelligenz (accessed 13.0.7.2024). *Der Bundesrat*.
- CAHAI (2021). A legal framework for AI systems (accessed 13.0.7.2024). *Ad hoc Committee on Artificial Intelligence of the Council of Europe*.
- Datenethikkommission der Bundesregierung (2019). Gutachten der Datenethikkommission der Bundesregierung (accessed 16.12.2019). *Berlin: Bundesministerium des Innern, für Bau und Heimat.*
- EU AI Act (2024). REGULATION (EU) 2024/… OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). OJ 2024 L .... (accessed 28.06.2024). *The european parliament*.
- 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. *Das europäische Parlament*.
- European Commission adoption consultation (2021). Artificial Intelligence Act (accessed 13.0.7.2024). *EDRi*.
- European Union Agency for Fundamental Rights (2019). Facial recognition technology: fundamental rights considerations in the context of law enforcement (accessed 08.02.2022). *European Union Agency for Fundamental Rights*.
- National New Generation Artificial Intelligence Governance Specialist Committee (2021). Ethical Norms for New Generation Artificial Intelligence Released (accessed 13.0.7.2024). *Center for security and emergin technology*.
- Repräsentantenhaus (2022). H.R.6580 - Algorithmic Accountability Act of 2022 (accessed 13.0.7.2024). *117th Congress (2021-2022) of the United States of America*.
- Richardson, Rashida et al. (2019). Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force (accessed 13.0.7.2024). *AI Now Institute*.
- Thouvenin, Florent et al (2021). Ein Rechtsrahmen für Künstliche Intelligenz. *Digital Society Initiative Universität Zürich.*
- Verbraucherzentrale Bundesverband (2021). Artificial Intelligence needs Real-World Regulation (accessed 13.0.7.2024). *Verbraucherzentrale Bundesverband*.
- White House Office of Science and Technology Policy (2022). Blueprint For An AI Bill Of Rights Making Automated Systems Work For The American People. *The White House*.

B.2 Further sources

- AI Now Insititute (2018). Automated Decision Systems - Examples of Government Use Cases (accessed 08.02.2022). *AI Now Insititute*.
- Assion, Simon (2014). Überwachung und Chilling Effect. *«Überwachung und Recht», Tagungsband zur Telemediucs Sommerkonferenz 2014, epubli GmbH, Berlin.*
- Bürgi, Urs (2022). Arbeitnehmerüberwachung (accessed 25.01.2022). *Law Media*. Author: Bürgi Nägeli Rechtsanwälte.
- Crawford, Kate (2021). Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence. *Yale University Press, New Haven, MA*
- Digitale Gesellschaft (2023). Datenschutz-Konzept der Digitalen Gesellschaft (accessed

24.06.2024). *Digitale Gesellschaft*. Presented at the Datenschutzfestival on the 03.11.2023.

- Eidgenössischer Datenschutz- und Öffentlichkeitsbeauftragter (EDÖB) (2023). Merkblatt zur Datenschutz-Folgenabschätzung (DSFA) nach den Art. 22 und 23 DSG (accessed 25.20.2023). *Eidgenössischer Datenschutz- und Öffentlichkeitsbeauftragter (EDÖB).*
- Ensign, Danielle, et al. (2018). Runaway Feedback Loops in Predictive Policing. *FAT 2018,* 160-171.
- Eser Davolio, M. et al. (2020). Auswirkungen der Falllastreduktion in der Sozialhilfe auf die Ablösequote und Fallkosten: Entschleunigung zahlt sich aus (accessedf 13.07.2024). *Schweizerische Zeitschrift für Soziale Arbeit*, 2019(25), pp. 31–51.
- Fanta, Alexander (2020). Datenschutzbehörde stoppt Jobcenter-Algorithmus (accessed 7.8.2022). Netzpolitik.org, 21.8.2020.
- FDA (2023). Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices | FDA (Zugriss am 28.10.2023). fda.gov.
- Frenkel, Sheera und Kang, Cecilia (2021). An Ugly Truth. *Harper Collins Publishers.*
- Gesichtserkennung-stoppen.ch (2021). https://www.gesichtserkennung-stoppen.ch/ (accessed 14.02.2022).
- Imhasly, Patrick (2021). Article Forschung an Raubgut. *NZZ am Sonntag*, 19.09.2021.
- NiederlandeNet (2020). RECHT: Gericht verbietet Betrugsbekämpfung mit Hilfe des Computerprogramms SyRi wegen Eingriffs in Privatsphäre (accessed 7.8.2022). *NiederlandeNet*, WWU Münster, 5.2.2020.
- OECD (2024). Explanatory memorandum on the updated OECD definition of an AI system. *OECD Artificial Intelligence Papers*, No. 8.
- Orwat, C. (2019). Diskriminierungsrisiken durch Verwendung von Algorithmen. I*nstitut für Technikfolgenabschätzung und Systemanalyse (ITAS)*, Karlsruher Institut für Technologie (KIT).
- O'Neil, Cathy (2016). Weapons of Math Destruction. *New York: Crown.*
- Penney, Jonathon (2016). Chilling Effects: Online Surveillance and Wikipedia Use (accessedf 13.07.2024). *Berkeley Technology Law Journal*, Vol. 31, No. 1, p. 117, 2016.
- Public Code, Public Money (n.d.). https://publiccode.eu/ (accessed 14.02.2022).
- Reclaim your Face (2021). https://reclaimyourface.eu/ (accessed 14.02.2022).
- Smith, Brad (2018). Facial recognition technology: The need for public regulation and corporate responsibility (accessed 08.02.2022), *blogs.microsoft.com*.
- Tufekci, Zeynep (2018). YouTube, the Great Radicalizer (accessed 8.2.2022), *The New York Times*, 10.3.2018.
- Willson, Michele (2017). Algorithms (and the) everyday. *Information, Communication & Society*, 20:1, 137-150, DOI: 10.1080/1369118X.2016.1200645.

B.3 Picture credits

- Front page picture: Foto - Robynne Hu, Unsplash license

C 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.
- **Artificial intelligence, AI:** These are systems or algorithms that can take over (complex) tasks from humans. Due to the controversy and breadth of the term, we refrain from providing a definition here and instead refer to Automated Decision-Making Systems in the context of this document.
- **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. 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. (according to AI Now, Richardson et al. 2019, p. 20)

- **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.
- **Foundation Model:** These are AI and ADM systems that can be used for a variety of known or currently unknown tasks due to the generality of their processing and output options.
- **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 section A.
- **Model (decision algorithms)**: An algorithm (see algorithm) that can recognize certain types of patterns and relationships in input data series. In doing so, the numbers of the input data series are calculated with other numbers (the coefficients of the model) according to the calculation rule (the architecture of the model).
- **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.

D Table of changes

Table of changes from version 1.0 to version 2.0 of this document.

Section (Version 1.0)	Changes
General	Version number increased to 2.0, additional employees added, former employees mentioned accordingly.
0. Executive summary	Newly added.
1. Introduction	Rewrite of the introduction.
2. Scope	Mention why we do not use the term AI. Note on nudging added.
3. Summary of the legal framework	Reversal of the burden of proof explained in more detail. Counterweight of freedom in self-classification vs. extended duties highlighted.
4. Societal relevance	Rewritten, added example of ADM systems in social services, mentioned and quoted the case of Winterthur.
5. A regulatory proposal for ADM systems	Title changed. Added explanation of legal entities. Inserted explanation of state funding for open-source libraries and tools.
6. Categorization	Section 6.1 rewritten and aligned with the protection goals, previous versions removed. Definition of “risk” integrated in 6.2. In 6.2 “guidelines from the EU Commission's AI Act” replaced by “concepts from the EU Commission's AI Act”. In subsection 6.3, a short section on ensuring legal certainty for companies (esp. through ADMS supervisory guidance notes) was added.
7. Due diligence and transparency obligations	The introductory paragraph has been formulated more clearly and the operator is mentioned as an example. Impact assessments are discussed. Reference to the documentation of the classification into a risk category and to risk management has been added. References to existing regulations for state actors have been added. Reference to guidelines for AI and evaluation has been added.
8. Control, measures, and sanctions	Mention of the current state of the revision of the ZPO in a footnote. Deletion of the comparison of ADMS supervision with FINMA. Reversal of the burden of proof explained in more detail. Chain of recourse mentioned in 8.1 and 8.2. Responsibility of the operator remains.
9. Suggestions for the future	Minor changes.
A Feedback loops	Deleted.
B Regulatory proposals for ADMS, artificial intelligence, and algorithms	Appendix B updated, renamed to appendix A. Title slightly adjusted.