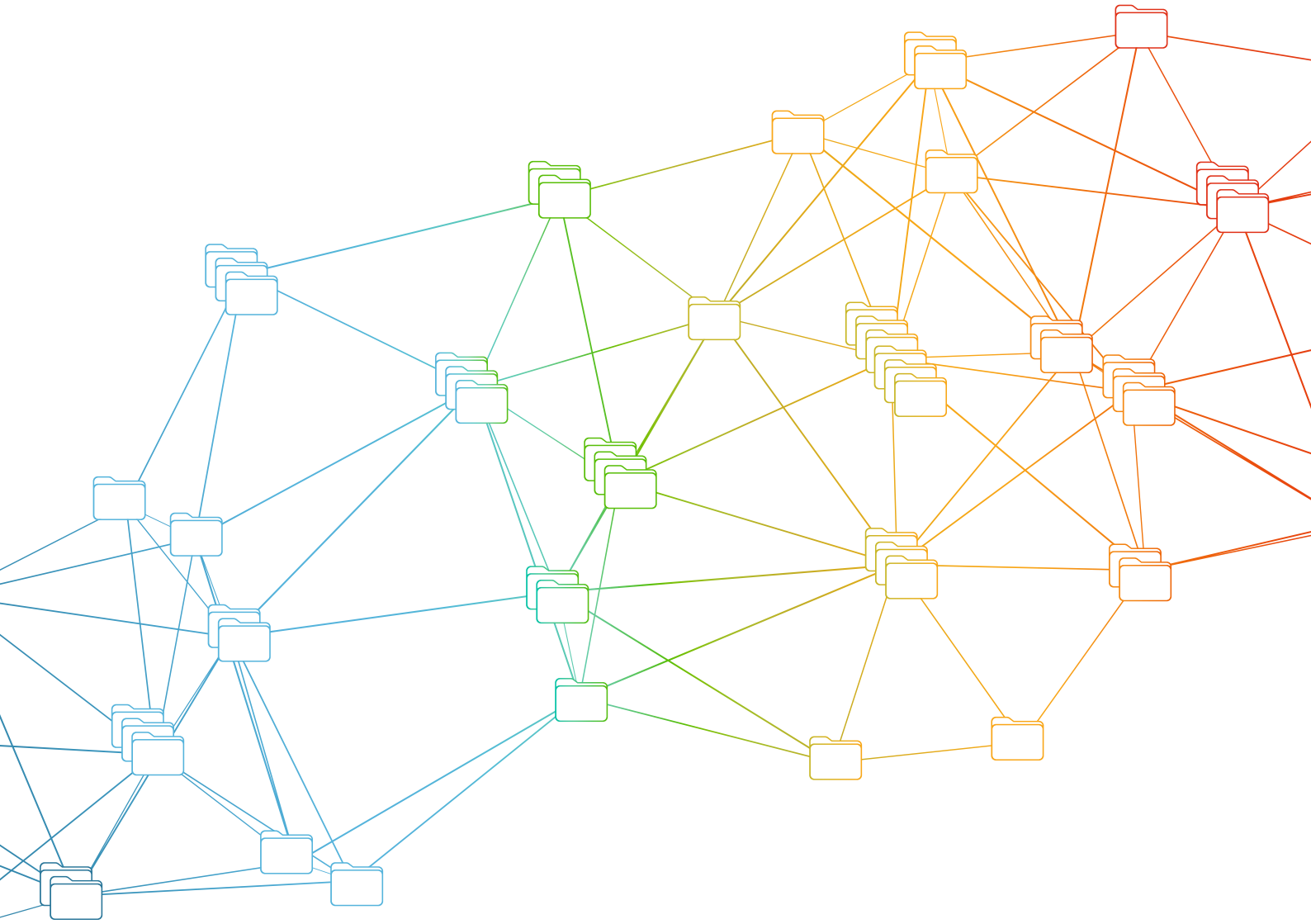
Automated Decision-Making Systems in the Public Sector

An Impact Assessment Tool for Public Authorities

Michele Loi, in collaboration with
Anna Mätzener, Angela Müller, and Matthias Spielkamp



ALGORITHM WATCH

ALGORITHM WATCH / CH

/ Executive Summary and Policy Recommendations

When deploying automated decision-making systems (ADMS) in the public sector, individual and societal trust in public authorities should be the ultimate benchmarks. Given the unique context in which public authorities act, the use of ADMS must be accompanied by a systematic assessment of potential ethical implications, ensuring transparency and accountability vis-à-vis those affected. In collaboration with the University of Basel, AlgorithmWatch conducted a study on the use of artificial intelligence in public administration on behalf of the Canton of Zurich, Switzerland.[1] As a result of this study, we developed a concrete and practicable impact assessment tool, ready to be implemented for the evaluation of specific ADMS by public authorities at different levels.

In the present report, we first outline the ethical foundation of our approach, from which specific questions for the evaluation of ADMS are derived. If the goal is to uphold the principles of harm prevention, autonomy, justice and fairness, and beneficence, there is no way around guaranteeing transparency, control, and accountability. The latter are not goals in themselves but unavoidable means toward guaranteeing an ethical use of ADMS in the public sector and for ensuring accountability. Based on that, we introduce a two-stage impact assessment procedure. It enables a triage of ADMS, indicating whether a system must be subject to additional transparency requirements. If this is the case, public authorities must provide a comprehensive transparency report. Lastly, we illustrate the use of the impact assessment tool by way of a fictional example.

/ Policy Recommendations

- When ADMS are deployed in the public sector, authorities should be obliged to systematically evaluate and make transparent the potential risks of their use. These risks cannot be determined in a generalized manner but only through a case-by-case analysis. Thus, it should be mandatory for public authorities to conduct an impact assessment prior to and during the deployment of any ADMS.

- The two-stage impact assessment procedure developed here provides a practicable and ready-to-use tool to generate transparency on these potential risks, based on seven underlying ethical principles. It enables a triage of ADMS, indicating whether a specific system must be subject to additional transparency requirements. If this is the case, public authorities must ensure that a comprehensive transparency report is provided, allowing for the evaluation of the system and its deployment over its entire life cycle. Transparency does not by itself ensure conformity with ethical requirements, but it is a necessary condition for achieving such conformity.

- There should be a public register for all ADMS deployed in the public sector. Next to information on its purpose, the underlying model, and the developers and deployers of the system, this register should contain the results of the impact assessment—that is, if applicable, the transparency report. In cases where there are legitimate reasons for restricting access to the transparency report, information should be provided on the fora to which it is disclosed.

1 Braun Binder, Nadja, et al., *Einsatz Künstlicher Intelligenz in der Verwaltung: rechtliche und ethische Fragen*, (Kanton Zürich; Universität Basel: February 2021). The final report [in German] is open access and available here: https://www.zh.ch/content/dam/zhweb/bilder-dokumente/themen/politik-staat/kanton/digitale-verwaltung-und-e-government/projekte_digitale_transformation/ki_einsatz_in_der_verwaltung_2021.pdf.

Contents

1 Introduction

The use of ADMS has arrived in the public sector. As a result of the most comprehensive mapping of ADMS use in sixteen European countries in our Automating Society Report, we now speak of the "automated society".[2] **In the years to come, the automation of decision-making procedures and services in public administrations is likely to increase exponentially. Citizens demand user-friendly services that are simple, easily accessible, and available 24/7. Administrations regard automation as a chance to accelerate efficiency, facilitate processes, and expedite mass and routine services. However, given the unique context in which public authorities act, the deployment of ADMS should be accompanied by a systematic evaluation of potential ethical implications, ensuring transparency and accountability vis-à-vis those affected. In collaboration with the University of Basel, AlgorithmWatch conducted a study on the use of artificial intelligence in public administration on behalf of the Canton of Zurich, Switzerland. As a result of this study, we developed a concrete and practicable impact assessment tool, ready to be implemented for the evaluation of specific ADMS by public authorities at different levels.**[3]

Chatbots facilitate the provision of information to citizens, the electronic processing of tax declarations speeds up assessments while saving resources, and the automated processing of citizen complaints can efficiently allocate different tasks to the corresponding public agency. Next to the above, ADMS are used for processing requests for social benefits, for detecting risks of welfare fraud, for profiling unemployed people, for predictive policing purposes by law enforcement authorities, or for assessing recidivism risks of parolees.

Doubtlessly, ADMS offer great potential for public administrations. **At the same time, it comes with substantial risks**—especially if such systems are not introduced and deployed in a careful manner. While these risks are certainly not limited to the **public sector** (but often mirror similar risks that pertain to the private sphere), the latter is of a special kind. Actions by public authorities are not only subject to different and unique legal preconditions, such as to the principles of legality or compliance with fundamental rights. They also take place in a unique setting where individuals do not have the freedom to choose the provider of services but are inescapably subject to a particular administration, according to rules of jurisdictional authority. In addition, decisions by public authorities often have consequential effects on individuals. When deploying ADMS in the public sector, this context must be considered—**individual and societal trust in public authorities should be the ultimate benchmarks for the automation of administrative procedures.**

2 Chiusi, Fabio; Fischer, Sarah; Kayser-Bril, Nicolas; Spielkamp, Matthias (eds.), *Automating Society Report 2020*, (AlgorithmWatch; Bertelsmann Stiftung: October 2020), https://automatingsociety.algorithmwatch.org.

3 For the full study, see Braun Binder, Nadja, et al., *Einsatz Künstlicher Intelligenz in der Verwaltung: rechtliche und ethische Fragen*, (Kanton Zürich; Universität Basel: February 2021). The final report [in German] is open access and available here: https://www.zh.ch/content/dam/zhweb/bilder-dokumente/themen/politik-staat/kanton/digitale-verwaltung-und-e-government/projekte_digitale_transformation/ki_einsatz_in_der_verwaltung_2021.pdf.

As a result, it is of prior importance not only to watch and unpack the use of ADMS in the public sector but also to contribute to a foundational **ethically oriented reflection** on the risks of this use, its impacts on individuals and society, and the requirements systems have to comply with. Yet, as worthy as this reflection is at the meta-level, **it does not suffice.**

In light of the consequential effects ADMS can have, public authorities should be obliged to assess the potential risks of any system they plan to deploy in a comprehensive, transparent, and systematic manner. Hence, in order to make a difference in practice, **ethical reflection must be translated into practicable and ready-to-use tools**, providing authorities with the means for conducting such an analysis. In what follows, **we provide a practical and concrete tool, enabling impact assessment of ADMS on a case-by-case basis and throughout their entire life cycles, ready for implementation by public authorities.** With this, we aim to contribute not only to the debate on the ethical use of ADMS but also to the provision of practical solutions—and thus, ultimately, to make a difference on the ground, that is, to ensure that these systems are used to the benefit of society as a whole, and not to its detriment.

If the goal is to evaluate the risks, stakeholders need **transparency**. Without any light shed on the functioning of concrete ADMS, their purpose, the actors involved, and their potentially harmful effects, they will remain black boxes—to the administration and its personnel, to those affected, and to society as a whole. **While transparency does not guarantee the ethical conformity of a system, it is a necessary precondition for ensuring such conformity and for enabling accountability.**

Thus, in what follows, we introduce a **two-stage evaluation procedure**, which can be used in order to detect the ethically relevant implications of a particular ADM system. It enables a **triage** of such systems, indicating whether a system must be subject to additional transparency requirements. If this is the case, public authorities must ensure the provision of a comprehensive **transparency report**. Thus, the approach developed here departs from the idea that ex ante risk assessments could be provided for certain categories of ADMS in a generalized manner. Rather, every ADMS must be subject to a case-by-case impact assessment.

The results of this impact assessment should be disclosed within a **public register**. Such registers should be mandatory for all ADMS used in the public sector and contain, in addition, information on the purpose of the system, its underlying model, and the actors involved in developing and deploying it. Such registers also enable independent external reviews by giving external researchers (academia, civil society, and journalists) access to relevant data on ADMS. This contributes to public interest research and thereby to an evidence-based debate on the automation of the public sector—a prerequisite for guaranteeing democratic control and accountability.

However, transparency does not always equal full public disclosure. In specific contexts, legitimate interests might speak against giving the public full access to transparency reports (such as, for example, the protection of personal data). However, in such cases, transparency must be provided vis-à-vis specific fora, e.g., the appropriate oversight institution. This, in turn, must be publicly communicated.

While ethical reflection on ADMS must eventually be translated into practicable tools, it is also true that the latter cannot be developed in a vacuum. They need to be based on **coherent foundational analyses and on substantial principles**, which must be **made explicit**. While these two levels are sometimes separated in the current debate on ADMS, we regard it as critical that they go hand in hand. Thus, before presenting the checklists mentioned above, the next section will justify and explain the ethical foundation of our approach, that is the seven theoretical principles this practical toolbox is built upon.

In section 3, the questions that flow from these principles are developed into two separate checklists. The *triage checklist* for ADMS (checklist 1) helps to determine which ethical transparency issues need to be addressed and documented prior to and during the implementation of the ADM project. The *transparency*

report checklist (checklist 2) then serves as a guide for compiling a detailed *transparency report.*

By means of the fictional example of a *Swiss COMPAS* risk evaluation system for criminal offenders, the application of the evaluation procedure is illustrated in section 4. Finally, a flow chart provides an overview of the approach as a whole (section 5).

2 Ethical Framework

2.1 Approach and Choice of Principles

The choice of the principles on which this approach is built flows from a review of a broad range of ethics guidelines, conducted as part of a larger study by AlgorithmWatch on the ethical use of ADMS in public administration.[4]

We will base our approach on the "Ethics Guidelines for Trustworthy AI" drafted by the Independent High-Level Expert Group on Artificial Intelligence, which was set up by the European Commission in 2019 (these are referred to as "EU Ethics Guidelines" in what follows).[5] However, these guidelines are a rough simplification of the values introduced by other guidelines and thus need to be complemented. We do not discuss the fundamental moral justifications of these principles; we assume, in the light of the significant convergence found in the *practical* content of many guidelines, that practical principles—regulatory principles—are all supported by overlapping fundamental moral concerns.[6] That is to say, different moral doctrines support the same regulatory principles, for example, the principle of ensuring human control of technology can be justified by appealing to autonomy or to the principle of respect of human dignity.[7] We, therefore, pick up one of the many possible formulations of a subset of these apparently common principles, while assuming that nothing of fundamental moral or political importance follows from this somewhat arbitrary choice.

The most comprehensive analysis of AI guidelines published so far[8] includes a list of eleven distinct principles, which constitute the common denominator of the guidelines under scrutiny. Of these eleven principles, one can distinguish ethical principles that overlap with the EU Ethics Guidelines (*non-maleficence or harm prevention*, *justice and impartiality [fairness]*, and *freedom and autonomy*); ethical principles not included in the EU Ethics Guidelines as such (*beneficence* and *respect for dignity*); and instrumental, technical or procedural requirements included in the EU Ethics Guidelines under the heading of "Key Requirements of Trustworthy AI" (*transparency* and *responsibility/accountability*). A further principle (*explicability*) is included in the EU Ethics Guidelines but (convincingly) considered an enabler of other principles, such as the implementation requirements.[9] The ethical principle of *beneficence* is not included in the EU Ethics Guidelines, but it forms part of both the summary referred to above[10] and

4 For the full review of ethics guidelines, see Braun Binder, et al., *Einsatz Künstlicher Intelligenz in der Verwaltung*, 65-66.

5 Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, "Ethics Guidelines for Trustworthy AI" (European Commission – Digital Single Market, April 8, 2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

6 John Rawls, *A Theory of Justice*, 2nd ed. (Cambridge, MA: Harvard University Press, 1999).

7 John Rawls, *Political Liberalism*, Expanded ed. (New York: Columbia University Press, 1996).

8 Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," Nature Machine Intelligence 1, no. 9 (September 2019): 389–99, https://doi.org/10.1038/s42256-019-0088-2.

9 Luciano Floridi and Josh Cowls, "A Unified Framework of Five Principles for AI in Society," Harvard Data Science Review 1, no. 1 (July 2019), https://doi.org/10.1162/99608f92.8cd550d1.

10 Jobin, et al., "The Global Landscape of AI Ethics Guidelines."

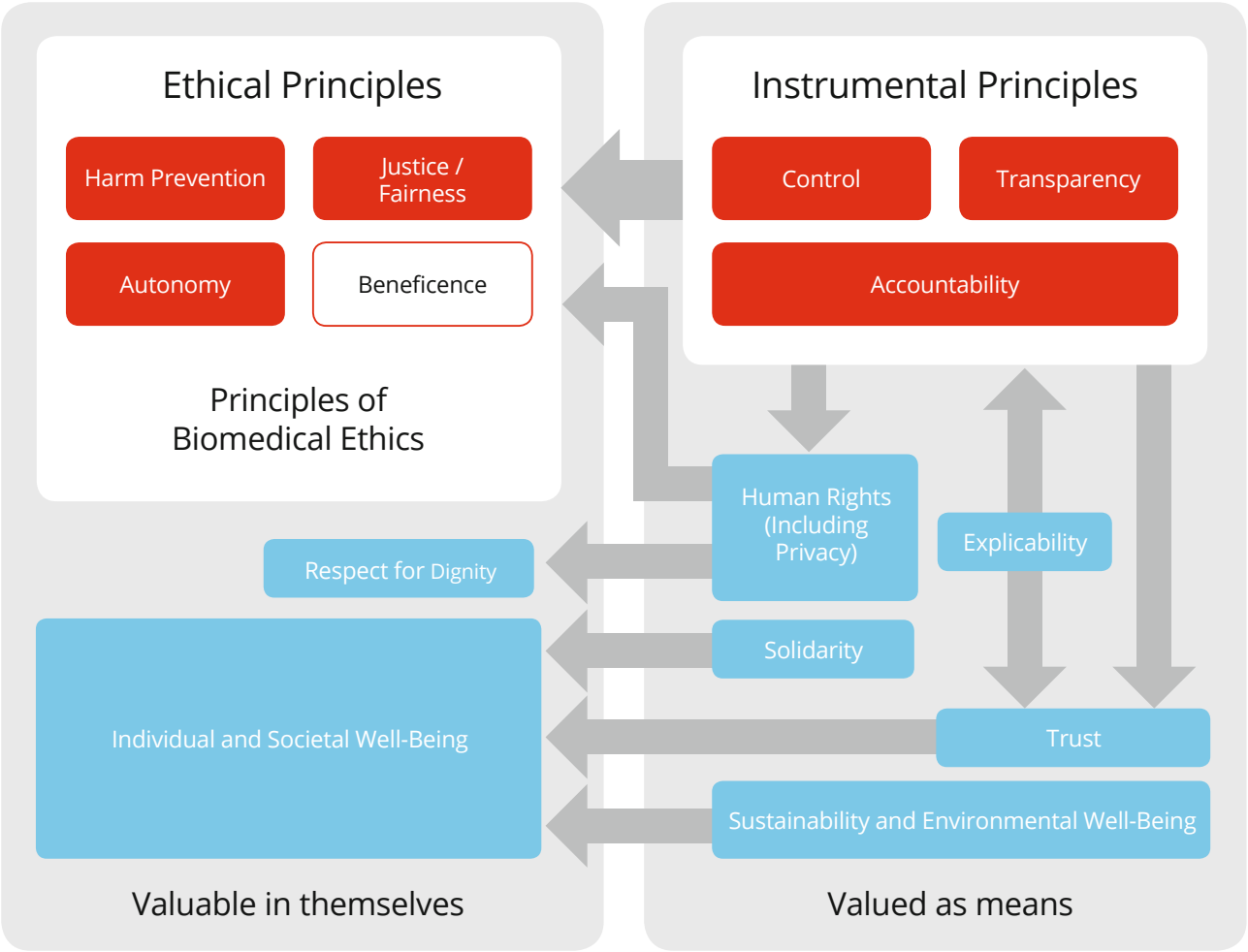


Fig. 1, the most important principles and values in ethical guidelines on AI (illustration by authors). In red, the principles considered in this framework; in blue, values and principles appearing in other guidelines. The arrow means "is required for". For simplicity, we include environmental well-being as a means to enable humans to flourish. Arguably, respect for the environment can also be regarded as a moral end in itself, quite apart from its impact on people. This is controversial.

the most widely used framework of ethical principles, those of biomedical ethics.[11]

Further principles found in many guidelines can be classified as belonging to three macro-categories: control, transparency, and accountability. Here, we will refer to control, transparency, and accountability as "instrumental principles".[12]

The analysis of eighteen other documents about the use of ADM in the public sector[13] reveals further ethical and instrumental principles compatible with this structure.

11 Tom L. Beauchamp and James F. Childress, Principles of Biomedical Ethics, 6. ed. (New York: Oxford University Press, 2008).

12 Michele Loi, Christoph Heitz, and Markus Christen, "A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data," in 2020 7th Swiss Conference on Data Science (SDS), 2020, 41–46, https://doi.org/10.1109/SDS49233.2020.00015; Michele Loi, "People Analytics Must Benefit the People. An Ethical Analysis of Data-Driven Algorithmic Systems in Human Resources Management" (AlgorithmWatch, March 2, 2020), https://algorithmwatch.org/wp-content/uploads/2020/03/AlgorithmWatch_AutoHR_Study_Ethics_Loi_2020.pdf. A similar value framework, including three of the four principles from the EU Ethics Guidelines (non-maleficence or harm prevention, justice and impartiality [fairness], and freedom and autonomy) plus beneficence, plus the three key procedural requirements of control, transparency, and accountability are used in the AlgorithmWatch report on the ethics of human analytics and algorithms in HR.

13 Braun Binder, et. al., *Einsatz Künstlicher Intelligenz in der Verwaltung*, 65-66.

As a result of this review, we focus on a framework of seven values:

- four ethical principles, namely respect for human autonomy, prevention of harm, justice or impartiality [fairness], and beneficence
- three instrumental principles, namely control, transparency, and accountability, summarizing technical, organizational, and prudential requirements[14]

In what follows, we present the ethical framework that forms the basis of our practical recommendations and checklists in more detail. For each of the seven values considered, we provide references to analogous concepts in existing guidelines for public sector applications of ADMS. Moreover, references to the checklist items of checklists 1 and 2, which are directly inspired by this analysis, are put into italics (e.g., *question 1.10* or *question 2.8*).

2.2 Seven Principles

2.2.1 Ethical Principles

/ Harm Prevention

This is the principle of "Do no harm". Civilian ADMS must not be designed so as to harm or deceive people and should be implemented in ways that minimize any negative outcomes.[15] First and foremost, and in its simplest form, avoiding harm means not inflicting pain and discomfort on people. In a broader sense, harm includes violations of privacy[16] (*question 1.1*) and violations of rights (*questions 1.4 and 1.5*), including human rights.[17] Avoiding harm is associated with promoting safety, security,[18] and sustainability, and more broadly with building technical and institutional safeguards.[19] These are often mentioned as elements of trust or trustworthy technology.[20] Harm avoidance also involves preventing and managing risk, which often requires ensuring that systems are reliable[21] and predictable.[22] Harm prevention includes ensuring broad environmental and social sustainability (*question 1.10*),[23] because non-sustainable practices will ultimately cause human harm and run counter to own interests. Moreover, the sustainability

14 Explicability (or explainability) is not considered an independent principle but a component of other instrumental principles, cf. Jobin, et al., "The Global Landscape of AI Ethics Guidelines"; Loi, et al., , "A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data."

15 D Dawson et al., "Artificial Intelligence: Australia's Ethics Framework," (Data61 CSIRO, 2019), https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf, cf. "Principle 2".

16 Dawson et al., "Artificial Intelligence: Australia's Ethics Framework," cf. "Principle 4".

17 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems" (April 8, 2020), https://rm.coe.int/09000016809e1154, cf. Principle 3.2; Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making" (February 5, 2019), https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592, cf. "Appendix B" (impact assessment levels).

18 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems"; Schweizerische Eidgenossenschaft – Der Bundesrat, "Leitlinien ‚Künstliche Intelligenz' für den Bund. Orientierungsrahmen für den Umgang mit Künstlicher Intelligenz in der Bundesverwaltung," (November 25, 2020), https://www.sbfi.admin.ch/dam/sbfi/de/dokumente/2020/11/leitlinie_ki.pdf.download.pdf/Leitlinien%20K%C3%BCnstliche%20Intelligenz%20-%20DE.pdf, cf. "Leitlinie 5"; Jan Engelmann and Michael Puntschuh, "KI im Behördeneinsatz: Erfahrungen und Empfehlungen" (December 2020), https://oeffentliche-it.de/documents/10181/14412/KI+im+Beh%C3%B6rdeneinsatz+-+Erfahrungen+und+Empfehlungen, cf. "5. Sicherheit."

19 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems"; Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. "5. Sicherheit".

20 David Leslie, *Understanding Artificial Intelligence Ethics and Safety* (London: The Alan Turing Institute, 2019), 6, https://doi.org/10.5281/zenodo.3240529; Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, "Ethics Guidelines for Trustworthy AI".

21 Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," (November 20, 2020), https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter.

22 Guidelines and laws that require a risk assessment may include: "the rights of individuals or communities, the health or well-being of individuals or communities, the economic interests of individuals, entities, or communities, the ongoing sustainability of an ecosystem," Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making".

23 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Principle 6.3".

of a socio-technical system grounded in reliable technology may be considered here.[24] This also covers ensuring cybersecurity (*question 1.2*), including the protection of the confidentiality, integrity, and availability of information.[25]

/ Justice and Fairness

A basic principle of justice is the idea that one ought to treat equals equally and unequals unequally. In other words, an action is just if it treats everyone equally who is equal with regard to the criterion that is morally relevant in a particular situation. Thus, students must get equal grades for equal performance in an exam (performance being the morally relevant criterion in this situation). They must not get equal grades for equal height (as height is morally irrelevant in this situation). Hence, justice refers to equality—but to a qualified notion of equality.

Applied to the present issue, this means that if ADMS treat people unequally who are actually equal with regard to the criterion that should be morally relevant for the particular decision the system contributes to—and the system does so on the basis of criteria in which they are unequal but that should not be considered relevant for this particular decision—this violates basic principles of justice. For example, if an ADMS deployed in the context of social welfare comes to a pattern of different decisions because of gender (which should not be morally relevant for this particular decision), as such, it treats male and female persons unequally, even though they are equal with regard to the criteria that should be relevant for the assignment of social welfare benefits. Hence, the decision is unjust.

However, this is compatible with using gender, in many situations, as a proxy for other relevant aspects. Gender information may justifiably be "grounds" for the decision in a model that predicts the need for social services, if it is among the criteria that best predict who needs welfare assistance the most.[26] Moreover, gender information (and other demographic group information) may be considered and processed when it is necessary to correct and prevent inequalities that reflect socially established biases against a specific gender or other demographic groups, given that this does not conflict with the requirements of positive law. Thus, the evaluation of the fairness of an algorithm is morally complex and does not boil down to ignoring certain variables.

The ethical goal of justice and fairness must be safeguarded by means of six ethical principles. The first is protection against unjust discrimination and unjustifiable bias.[27] This dimension of fairness applies to different elements of the data pipeline:

If the ADMS processes social or demographic data, it should be designed so as to meet a minimum level of discriminatory non-harm. To achieve this, one should:

— use only fair and equitable datasets (data fairness),

24 Government Digital Service and Office for Artificial Intelligence, UK, "A Guide to Using Artificial Intelligence in the Public Sector/ Understanding Artificial Intelligence Ethics and Safety," (June 10, 2019), https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety.

25 Council of Europe, "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment" (April 8, 2020), https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c, cf. "Principle of Quality and Security"; Dawson et al., "Artificial Intelligence: Australia's Ethics Framework," cf. "Principle 4"; Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "6.3.7"; Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. "6. Datenhaltung und -qualität".

26 There is an important ambiguity in ordinary language when people refer to a feature, e.g., gender, as the "grounds" of a decision. Grounds can be epistemic or moral. If an algorithm uses gender to identify the neediest people, gender is the grounds of the decision only from the epistemic point of view: it is, at most, defeasible evidence that the individual is needy. It is not treated as being in itself a distinct moral reason to provide welfare assistance.

27 Council of Europe, "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment," cf. "Principle of Non-discrimination"; Dawson et al., "Artificial Intelligence: Australia's Ethics Framework," cf. "Principle 5"; Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," cf. "Fairness and Justice"; Leslie, "Understanding Artificial Intelligence Ethics and Safety"; Government Digital Service and Office for Artificial Intelligence, UK, "A Guide to Using Artificial Intelligence in the Public Sector"; Automated Decision Systems Task Force, "New York City Automated Decision Systems Task Force Report" (November 2019), https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf.

- include reasonable features, processes, and analytical structures in the model architecture (design fairness),
- prevent the system from having any discriminatory impact (outcome fairness),
- implement the system in an unbiased way (implementation fairness).[28]

Avoiding all biases may be impossible. Biases may be a result of inadequate data for training the model,[29] but finding representative data for training purposes may be hard or even impossible. Among other reasons, this is because of privacy protections and (paradoxically) because of anti-discrimination laws, which may put limits on collecting and processing data about, for example, race, gender, ethnic group, or religious status.

But even when data sets are fully adequate in terms of representing society in all its subdivisions and facets, decisions based on statistical generalizations may still be considered unjustly biased. For example, if they favor individuals with features that correspond to those from more privileged backgrounds (if this advantage for one group is not an unavoidable collateral effect of other necessary and morally justified features of the algorithm). This is especially the case if data are representative of different sub-populations: It raises the probability that some data (or combinations thereof) act as a proxy of race, gender, religion, etc. Even if data that explicitly refers to categories protected by anti-discrimination laws are absent from the mix, any efficient machine learning algorithm will typically learn to recognize proxies.

This is all the more predictable the more society is unequal. For example, in a society where most men have higher incomes than most women, knowing the gender, or anything that correlates with it, well enough, will help the algorithm to make fewer errors in identifying future high-income earners capable of sustaining a large mortgage. Regrettably, such inferences improve algorithmic accuracy and efficiency (as measured by the benchmark of past technology) *in the aggregate,* even if they may result in a large number of wrong predictions about individual cases. They may still improve efficiency even if society is changing and the income level of women is increasing and, as a result, the algorithm systematically overestimates the financial advantage of being male (which may remain unknown unless someone collects the data necessary to determine it). This is why all algorithmic inferences based on techniques of statistical learning from data about humans in an already unequal society are potentially morally problematic in terms of bias and indirect discrimination. In the checklist, these algorithms will thus require special attention (*question 1.13*).

Moreover, the dimensions of outcome fairness and application fairness are especially important when algorithms influence competitions in the political and the economic sphere (*questions 1.11* and *1.12*). Competitive processes are hugely important because our society relies on fair competition in the market and politics as the procedure to decide how to distribute social resources and opportunities in society in a way that can be considered procedurally just overall.[30]

The second dimension of fairness requirements reaches beyond ethics and refers to legality, ensuring that algorithms do not violate existing legal norms, including legal rights.[31]

Third, justice and fairness include respect for all rights, whether recognized in positive law or not, including human rights and moral rights.[32]

---

28 Government Digital Service and Office for Artificial Intelligence, UK, "A Guide to Using Artificial Intelligence in the Public Sector".

29 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Testing on Personal Data".

30 John Rawls, *A Theory of Justice*, 2nd ed. (Cambridge, MA: Harvard University Press, 1999).

31 Dawson et al., "Artificial Intelligence: Australia's Ethics Framework," cf. "Principle 3"; Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," cf. "Fairness and Justice".

32 Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," cf. "Fairness and Justice".

Fourth, it includes the values of equality,[33] inclusion, and solidarity (although this set of values may be a more contested and highly contextually variable). The requirement of inclusion, covering elements of public participation, appears to be more strongly affirmed in guidelines addressing the public sector[34] compared to guidelines addressing the scientific community or private companies, which tend to have been published earlier.

Fifth, the principle of justice and fairness includes compensation and remedies (*questions 1.7, 1.8, 1.9*) in case the violation of a right can be proven.[35]

Sixth, it refers to the issue of procedural regularity. Some AI-based ADMS continuously update their models based on new data in order to improve in achieving whatever goal they have been programmed to achieve.[36] One of the collateral effects of such continuously updating models is that they can produce different outputs from the same inputs, if the same input is processed before or after the update of the model. As a result, two individuals with the same characteristics (inputs) may receive a different decision (output), depending on when the algorithm processes their personal data (before or after a model update).[37] This may violate the rights of individuals to equal treatment by the law (*question 1.14*).

/ Autonomy

Promoting autonomy means enabling individuals to make decisions about their lives on their own, which are not imposed upon them or manipulated by others. It requires having a range of genuine options available from which one can choose. A decision taken on the basis of inadequate information or deception is not considered autonomous. The value of human autonomy is mostly related to the procedural requirement of transparency, which will be addressed in the second part of the present framework. Transparency involves providing sufficient, accurate information and avoiding deception of individuals interacting with algorithms, thereby enabling autonomous choices.

The most familiar (and perhaps least valued) exercise of decisional autonomy in the digital realm pertains to personal data. Individuals must be informed about what can happen to their personal data so that they can give their consent for their data to be used and processed in one way or another,[38] especially when it comes to experimental technologies.[39]

Another aspect of autonomy is the "ability to question and change unfair, biased or discriminatory systems",[40] which citizens only have if they obtain "understandable and accurate information about the technological, algorithmic and artificial intelligence

33 Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," cf. "Fairness and Justice".

34 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems."; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Barriers, Advancement of Public Benefits"; Dataethical Thinkdotank, "White Paper: Data Ethics in Public Procurement," https://dataethics.eu/publicprocurement/, cf. "Universal Design"; Cities for Digital Rights, "Declaration of Cities Coalition for Digital Rights" (July 15, 2020), https://citiesfordigitalrights.org/declaration, cf. "Participatory Democracy, Diversity and Inclusion"; Schweizerische Eidgenossenschaft – Der Bundesrat, "Leitlinien, 'Künstliche Intelligenz' für den Bund; cf. Leitlinie 7.

35 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Effective Remedies"; Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "Principle 6.4".

36 Note that such adaptability could mitigate the fairness problem highlighted above, namely, that of predictions reflecting obsolete social patterns from the past, which are needlessly harmful to social groups whose conditions are continuously improving. But continuous learning and model optimization raises its own fairness concerns. This is not surprising: if fairness involves distinct dimensions, it is possible that there will be dilemmatic choices when the designer must choose a solution that may be better with regards to one fairness aspect, but not with regards to a different one, and no solution exists that is optimal for both.

37 Joshua A. Kroll et al., "Accountable Algorithms," University of Pennsylvania Law Review 165, no. 3 (2017): 633; Michele Loi, Andrea Ferrario, and Eleonora Viganò, "Transparency as Design Publicity: Explaining and Justifying Inscrutable Algorithms," Ethics and Information Technology (2020), https://doi.org/10.1007/s10676-020-09564-w.

38 Cities for Digital Rights, "Declaration of Cities Coalition for Digital Rights," cf. "Privacy, Data Protection and Security"; World Economic Forum, "AI Government Procurement Guidelines" (June 2020), http://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_AI_Government_Procurement_Guidelines_2020.pdf.

39 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Computational Experimentation".

40 Cities for Digital Rights, "Declaration of Cities Coalition for Digital Rights," cf. "Transparency, Accountability, and Non-discrimination of Data, Content and Algorithms".

systems that impact their lives".[41] Contestation can be considered both as valuable in itself as an element of human autonomy and as instrumentally valuable, as a form of control over algorithms and as a way to promote accountability (the latter aspect will be considered in the second part of the present framework).

A further aspect of autonomy is the possibility to choose which digital services to use, or to avoid using them altogether,[42] especially if these are experimental services (*question 1.6*).[43]

Autonomy also has a collective dimension: it involves the ability of citizens as a group to take common decisions about their collective destiny as a community. This collective dimension of autonomy is not very common in AI guidelines,[44] but it seems very important for public digital infrastructure, such as that of smart cities.[45]

This collective dimension of autonomy is again grounded in individual autonomy, stressing the value of participation. It underlines the importance of being able—as an individual—to participate in collective decision-making processes of the society and communities of which one is a member.

The ethical requirement to respect fundamental rights[46] implicitly furthers autonomy, because many such rights (which are, typically, human rights that are constitutionally protected in democracies) are based on the idea of protecting human autonomy. For example, negative rights such as those to freedom of expression or freedom of religion protect individual autonomy in the public and political sphere as well as in the sphere of individual and collective expression. Positive rights such as the right to healthcare and education protect autonomy by ensuring that individuals have the means they need to live autonomous lives in dignity. Hence, the design and implementation of ADMS in socially sustainable ways is a means of promoting autonomy.[47]

Finally, in the realm of AI ethics, autonomy may refer to the idea of the AI-based ADMS being "under user control".[48] That is to say, algorithms should be used to assist human decision-making and not to replace it altogether. Most plausibly, this should be interpreted as a constrained principle. Philosophically speaking, it seems that automation can be (and has been) enhancing autonomy rather than reducing it, to the extent that automating routine tasks has been instrumental in providing humans with more time and resources to deal with more intellectually challenging, creative, or emotionally rewarding tasks.[49] The problematic implications with regard to autonomy stem from those forms of AI-based ADMS that are tasked with automating the more cognitively challenging types

41 Cities for Digital Rights, "Declaration of Cities Coalition for Digital Rights," cf. "Transparency, Accountability, and Non-discrimination of Data, Content and Algorithms".

42 Cities for Digital Rights, Declaration of Cities Coalition for Digital Rights," cf. "Open and Ethical Digital Service Standards".

43 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Computational Experimentation".

44 It was mentioned in only one guideline in the review of Jobin, et al., "The Global Landscape of AI Ethics Guidelines."

45 "Everyone should have [...] the ability collectively to engage with the city through open, participatory and transparent digital processes", in Cities for Digital Rights, Declaration of Cities Coalition for Digital Rights," cf. "Participatory Democracy, Diversity and Inclusion".

46 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Principle of Respect for Fundamental Rights".

47 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Human-centric and Sustainable Innovation"; Dataethical Thinkdotank, "White Paper," cf. "Human Primacy".

48 Council of Europe, "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment," cf. "Principle 'Under User Control'".

49 Michele Loi, "Technological Unemployment and Human Disenhancement," Ethics and Information Technology 17, no. 65 (2015): 1–10, https://doi.org/10.1007/s10676-015-9375-8; John Danaher, "The Threat of Algocracy: Reality, Resistance and Accommodation," Philosophy & Technology 29, no. 3 (2016): 245–268, https://doi.org/10.1007/s13347-015-0211-1; John Danaher, "Will Life Be Worth Living in a World Without Work? Technological Unemployment and the Meaning of Life," Science and Engineering Ethics 23, no.1 (2016): 41–64, https://doi.org/10.1007/s11948-016-9770-5.

of human activities, with humans taking orders from machines rather than giving orders to them.[50] Thus, autonomy as an ethical value is potentially at stake in all those automation projects where ADM is meant to replace human judgment (*question 1.15*) as well as with regard to all those forms of automation in which it is unclear whether the user sufficiently understands the decisions of the AI-based ADMS so that the latter can assist rather than replace her autonomy (*questions 1.16* and *1.17*). Furthermore, the public may lose control—and therefore autonomy—over its processes and decisions if it relies on an infrastructure that is entirely owned and made inaccessible by third parties *(question 1.18)*. This is an emerging concern, which was not identifiable in previous reviews[51] but it is one that seems quite important in relation to ADMS deployed in the public sector. [52]

/ Beneficence

Beneficence is arguably the fundamental principle of ethics that is least commonly found in AI guidelines. A plausible reason for this lack of attention is that most stakeholders take it for granted that ADM can produce benefits of some sort. Efficiency is often mentioned as a reason to use ADMS, the idea being that the same service can be provided to the same number of people while using less resources, or that existing services can be improved (e.g., made more accurate or furnished with extra features) while remaining affordable and thus accessible to most. Yet, the ability of ADM to do good is ethically fundamental: a framework for ethical ADM risks being heavily focused on harm prevention and less on the generation of benefits. Such a framework will tend to oppose the introduction of ADMS in most contexts because innovation as such generates risks. If one does not consider the potential benefit of innovation, there is, at first sight, no reason to take any risk. However, extreme risk avoidance is not always reasonable; rather, the risks inherent to innovation ought to be managed. For example, as mentioned in the last section, ADMS have the potential to enhance human autonomy when they are used to automate processes in a way that liberates human resources that can then be invested elsewhere. Fortunately, some guidelines addressing the public sector mention beneficence, at least implicitly by referring to the benefits of innovation,[53] for example: "Generates net-benefits. The AI system must generate benefits for people that are greater than the costs."[54] Beneficence is also implicitly addressed by guidelines that indicate the promotion of human well-being as the overall goal of such innovation.[55] Guidelines specifically addressing the public sector emphasize the idea that the benefit of ADM ought to be a public benefit.[56]

Beneficence is not mentioned explicitly in the checklist provided below. However, the concept of a net benefit is implicitly referred to in our transparency deliverable (Checklist 2), in particular in *items 2.1, 2.16, and 2.17*, which ask for explanations why the introduction of ADMS is useful and what evidence one can provide for this. Moreover, the issue of public good harm (*question 1.10*) leads (according to the proposed checklist algorithm) to a transparent stakeholder analysis (*question 2.8*) and a transparent presentation of the criticism that external stakeholders (*question 2.20*) could raise with regard to the cost-benefit analysis conducted by the public administration.

50 Loi, "Technological Unemployment and Human Disenhancement"; Danaher, "The Threat of Algocracy"; Danaher, "Will Life Be Worth Living in a World Without Work?".

51 Jobin, et al., "The Global Landscape of AI Ethics Guidelines".

52 Cities for Digital Rights, "Declaration of Cities Coalition for Digital Rights," cf. "Participatory Democracy, Diversity and Inclusion" and "Open and Dthical Digital Service Standards"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Infrastructure".

53 Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. „2 Interne Veränderung" and „3. Innovationsmarker".

54 Dawson, et al., "Artificial Intelligence: Australia's Ethics Framework".

55 Leslie, "Understanding Artificial Intelligence Ethics and Safety."; Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," cf. "Well-Being".

56 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Advancement of Public Benefit", "Rights-promoting Technology".

2.2.2 Instrumental and Prudential Principles

/ Control

The procedural requirement of control results from the analysis of twenty guidelines focusing on action types rather than action goals.[57] In the guidelines initially analyzed, it did not appear as a distinct requirement but rather as an element—a common set of actions—listed equally often under the rubric of transparency as well as under the rubric of accountability. Indeed, control includes common activities that are necessary for both transparency and accountability: one cannot be transparent about processes or outcomes of which one is not aware. Taking responsibility in the positive, forward-looking sense of the term (not in the sense of retrospective blameworthiness or liability) involves exercising control over processes so that they lead to the intended outcomes. Control includes all efforts that are necessary to confer robustness to all goal-related activities: from achieving the goal of the ADMS (as foreseen by the user—whatever that goal is—and realizing it) to ensuring that other ethical goals (harm avoidance, fairness, and autonomy) are also advanced. Control appears morally neutral because its ethical value is purely instrumental and contingent. A set of practices giving terrorists better control over ADMS used to coordinate drone attacks is useful but not ethically valuable, because the goal it enables terrorists to pursue is, as such, evil. Yet, when the agent's goal is ethically positive, the absence of control is ethically bad—and not merely neutral—because good intentions without control may fail to achieve the good they purport to achieve, or they can even unintentionally have harmful effects.

It is unsurprising—because it is such a wide-purpose means—control is the single most common procedural requirement found in AI ethics guidelines.

First of all, control includes the activity of documenting processes and outcomes as well as the recording,[58] testing,[59] and monitoring[60] that produces the data of what needs to be documented.[61] Documentation is distinct from transparency because it can be paired with internal communications, e.g., between collaborators within a data science team or an entire company,[62] which is compatible with lack of transparency to the user and the broader public.

Second, control includes measuring, assessing, evaluating,[63] and defining standards[64] and policies for all ADM-related processes and outcomes. It, therefore, includes the preliminary study necessary to generate meaningful standards and measures, that is to say, those that lead to an authentic understanding and knowledge of the processes and outcomes being assessed, which are both products and aspects of

57 Loi, et al., "A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data".

58 Dataethical Thinkdotank, "White Paper," cf. "Traceability".

59 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Testing".

60 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Interaction of Systems".

61 Schweizerische Eidgenossenschaft – Der Bundesrat., "Leitlinien ‚Künstliche Intelligenz' für den Bund," cf. „Leitlinie 3"; Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. „7. Wirkungsmonitoring".

62 Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. "8. Nachvollziehbarkeit".

63 Dillon Reisman, et al., "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability," AI Now Institute (2018): 1–22; AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI," n.d., https://assets.mofoprod.net/network/documents/Using_procurement_instruments_to_ensure_trustworthy_AI.pdf, cf. "Key Elements of a Public Agency Algorithmic Impact Assessment, #1"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Ongoing Review", "Evaluation of Datasets and System Externalities", "Testing on Personal Data"; World Economic Forum, "AI Government Procurement Guidelines," cf. "Data Quality"; Automated Decision Systems Task Force, "New York City Automated Decision Systems Task Force Report," cf. "Impact Determination".

64 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Standards"; Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. "1. Zielorientierung".

control over ADM. Understanding and knowledge of ADM require some form of explanation[65] of what these systems do. Thus, the goals of explainable AI and the value of explicability[66] found in the EU Ethics Guidelines are one facet (perhaps the most important facet) of the procedural requirement of control. Evaluation and assessment of an ADMS require justification[67] for design choices and, when unavoidable, its errors, biases, and trade-offs with other moral goals.

Third, control includes those social activities that are necessary to ensure that one's study of processes and outcomes is adequately complete and that it does not exclude relevant perspectives. This includes activities such as training[68] and enhancing internal expertise,[69] expert review,[70] and even diversity in the workforce,[71] and transparency as public debate,[72] when advocated as a tool for improving an organization's understanding of the social implications of ADMS.

Fourth, control includes risk-mitigation measures, such as building backups and contingency plans,[73] limiting and separating, blocking, and interrupting[74] processes, making room for human intervention,[75] predicting and preventing risks, prohibiting harmful or risky practices,[76] contesting[77] and correcting[78] errors that are made. The importance assigned to risk assessment and management[79] in the guidelines we have analyzed can hardly by overstated.

Fifth, and of special importance for the use of ADM in the public sector, control includes ownership,

---

65 Brent Mittelstadt, Chris Russell, and Sandra Wachter, "Explaining Explanations in AI," in Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19 (New York: Association for Computing Machinery, 2019), 279–88, https://doi.org/10.1145/3287560.3287574; Automated Decision Systems Task Force, "New York City Automated Decision Systems Task Force Report," cf. „Explanation"; Cf. Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. „8. Nachvollziehbarkeit".

66 Dataethical Thinkdotank, "White Paper," cf. "Explainability"; cf. Floridi and Cowls, "A Unified Framework of Five Principles for AI in Society"; Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand." Cf. "Transparency".

67 Leslie, "Understanding Artificial Intelligence Ethics and Safety"; Government Digital Service and Office for Artificial Intelligence, UK, "A Guide to Using Artificial Intelligence in the Public Sector," cf. "Transparency"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Testing"; Loi, Ferrario, and Viganò, "Transparency as Design Publicity".

68 Council of Europe, "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment," cf. "Personnel Management"; Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. „9. Akzeptanz".

69 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI," cf. "Executive Summary"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems." cf. "Independent Research" and "Rights-promoting Technology".

70 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI," cf. "Key Elements Of A Public Agency Algorithmic Impact Assessment, #2"; Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "Appendix C"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems." cf. "Consultation and Adequate Oversight" and "Expertise and Oversight".

71 Council of Europe, "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment," cf. cf. "Principle of Quality and Security" ; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Personnel management".

72 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Public Debate", also considered under "Transparency".

73 Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "Appendix C".

74 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems." cf. "Follow up", "Consultation and Adequate Oversight".

75 Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," cf. "Human Oversight and Adaptability"; Council of Europe, "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment," cf. "Principle 'Under User Control'".

76 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Consultation and Adequate Oversight" and "Follow up".

77 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Barriers" and "Effective Remedies".

78 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Consultation and Adequate Oversight" and "Effective Remedies".

79 World Economic Forum, "AI Government Procurement Guidelines," cf. "Key Variables to Consider in a Risk Assessment"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Human Rights Impact Assessment"; Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," cf. "Assessing Likelihood and Impact", "Human Oversight and Accountability", "Reliability, Security and Privacy"; Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "Algorithmic Impact Assessment".

knowledge, and effective control of some key infrastructure,[80] e.g., of the data assets and the machine learning algorithms to learn from them, which is essential for shaping, better knowing, and more tightly controlling the ADMS in use.

The guideline developed here does not provide a risk assessment tool that aims to quantify risk. The tools that do so among those examined, used by the governments of Canada[81] and New Zealand,[82] or recommended by the World Economic Forum Guidelines,[83] all have features that we evaluate as problematic—especially if intended to ground a legal requirement that ought to be sanctioned when violated. The New Zealand risk tool relies fully on subjective assessments of risk, asking its users to assess whether the likelihood of risk is "occasional", "improbable", or "probable" and whether the impact is "low", "moderate", or "high", but it fails to specify concrete, objective criteria on which to ground such judgments. Words such as "non-serious", "moderate", "widespread", or "serious" are highly contextual and vague. In a context in which classifying an application as having more serious consequences implies a requirement to apply costly documentation and management requirements, it is to be expected that the vagueness of language will lead the users to downplay the description of the risk (if they have to bear the consequences). Conversely, suppose that people in charge of evaluating risk do not bear the burden of the additional costs deriving from the consequential additional procedures. They may then be biased in favor of assigning a higher risk-level, making other people's work needlessly complicated, but protecting themselves from any liability. Moreover, the impact criterion considers both spread (the number of people affected) and severity (how serious the damage is) and it leaves unexplained how these dimensions comparatively contribute to risk.[84]

A similar criticism can be levied against the "Impact Assessment Levels" of the Canadian "Directive on Automated Decision-Making",[85] since it offers no criteria to distinguish "little", "moderate", "high" and "very high" impact. Moreover, the word "often" in the expression "will often lead to" is amenable to a wide range of conflicting interpretations.

The World Economic Forum Guidelines say, "it is important to consider factors such as how many people are impacted". However, it is unclear whether this is compatible with the anti-utilitarian principles guiding the laws of many EU countries. This type of risk indicator may suggest that a low-risk classification (and, as a result, a low level of control) for risky algorithms, able to do significant harm to individuals, is ethically permissible—as long as the majority substantially benefits from their introduction and only a small group of individuals is negatively affected by them. Without additional refinements, the principle may suggest (or be misunderstood as) a utilitarian calculus that runs counter to rights-based, justice-based political principles, which aim to protect the few and their fundamental rights, even from measures that would benefit the majority.[86]

Such a utilitarian calculus would also be in conflict with one of the Council of Europe's recommendations, which, when addressing the "responsibilities of private sector actors with respect to human rights and fundamental freedoms in the context of algorithmic systems", adopt a clearly anti-utilitarian ethical view

80 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Infrastructure" and "Interaction of Systems".

81 Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "Appendix B".

82 Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," see "Assessing Likelihood and Impact".

83 World Economic Forum, "AI Government Procurement Guidelines".

84 Suppose that the impact of a system has aptly been described as "widespread" and severity aptly described as "moderate". Based on the indication given, this could either be classified as "moderate risk" (because the severity level is moderate) or as "high risk" (because it is widespread), with significantly different consequences. The explanation provided in the risk matrix makes it impossible to know how to disentangle this.

85 Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making".

86 Ronald Dworkin, *Taking Rights Seriously* (Cambridge MA: Harvard University Press, 1977). Cf. "Each person possesses an inviolability founded on justice that even the welfare of society as a whole cannot override," John Rawls, *A Theory of Justice*, 2nd ed. (Cambridge, MA: Harvard University Press, 1999), 3.

by explicitly stating[87] that responsibilities to respect human rights apply "regardless of their size" (which may be considered a proxy for how many people are impacted).[88]

The approach pursued here also emphasizes the significance of control, which it particularly implements through documentation. Writing a transparency report is not only—or even mainly—intended to enable external control and verification, but also—and mainly—in order to oblige the administration to engage in clearly structured and targeted documentation, measurement, and assessment tasks. *Questions 2.1* and *2.2* require documenting the goals of the system; *questions 2.13* and *2.15* require documenting the responsibilities and human control structure available to control the system; *questions 2.7 and 2.8* and *questions 2.10, 2.12, 2.16, 2.17, 2.18, and 2.19* require documenting definitions, standards, testing, measurements, and assessments of performance, privacy, fairness, and stakeholders involved; *questions 2.9, 2.10, 2.14, 2.15, and 2.20* require documenting risk management structures, including monitoring, feedback, error correction, and cybersecurity as well as their results.

Our checklist, therefore, does not require the administration to quantify the degree of risk prior to dealing with it. However, it still works as a risk reduction tool, because it activates specific documentation tasks under the condition of specific risk signals being detected by the person answering the checklist by means of a hermeneutic methodology. The very act of asking and answering the questions in one checklist ought to make certain risks more visible, and the act of documenting serves as a reminder that risk safeguards must be in place. Thus, the higher the number of risk signals, the longer and more structured the documentation and transparency requirements and the burden for the administration in controlling the respective ADMS. It is conceived as a checklist for risk-response rather than for risk assessment.[89]

/ Transparency

Here, transparency is exclusively understood as the production and communication of information to parties outside the institution designing or implementing an ADMS. This includes auditors, external experts, journalists, politicians, individuals with responsibilities in other branches of the administration, and the broader public. Communication among the members of the data science team, or between a data science team and the CEO of the data analytics company to which they report, or between a private company and the administrative office in charge of procurement is considered internal, as all these parties can be taken to pursue the same goal with respect to the use and application of an ADMS.

There are at least four main theories about why transparency is instrumentally valuable from an ethical point of view.[90] First of all, there is the view that "sunlight is the best disinfectant", to cite Justice Louis Brandeis. According to this view, transparency promotes accountability, which, in turn, prevents at least the worst unethical behavior from occurring. Second, there is the idea that transparency contributes to the quality of the technology because it enables the crowd-sourcing of expert opinion and feedback of concerned citizens, which leads to better scrutiny of the technology, which in turn makes it more trustworthy. A third approach regards transparency as enabling end-users of a technology—or people who may be affected by it—to make an informed choice of whether to use it. Fourth, there is the view that

87 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf "1.2. Scale of Measures".

88 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Appendix B".

89 There are similarities, though; for example, checklist items 1.6 - 1.9 mirror some of the high-risk indicators of the Canada risk-assessment tool (Annex B) and elicit a requirement to answer specific sections of the transparency report.

90 Tal Z. Zarsky, "Transparent Predictions," U. Ill. L. Rev., no. 4 (2013): 1503; Paul B. de Laat, "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?," Philosophy & Technology 31, no. 4 (2017): 525–541, https://doi.org/10.1007/s13347-017-0293-z; Heike Felzmann, et al., "Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns," Big Data & Society 6, no. 1 (2019): 1–14, https://doi.org/10.1177/2053951719860542; Loi, et al., "Transparency as Design Publicity".

transparency enables public debate, which is necessary for the democratic legitimacy of technological solutions—which is especially important when the implementation of technology is not value-neutral. All these four approaches to transparency are reflected in the guidelines examined here, and they are all generally taken to contribute to trust in the technology.

The plausibility, strength, and scope of each of these theories are, however, debated and debatable.[91] The accountability theory may not be valid in all contexts but only in those that—for one reason or another—grab public attention. The crowd-sourcing theory may only support weaker forms of transparency, subjecting technology to the scrutiny of a restricted and selected group of experts. The autonomy theory may not take into proper account the limited cognitive resources of individuals who need to choose a technology, and it does not apply to cases in which individuals are not offered a choice whether or not to be affected by the technology. Lastly, the public debate theory may fail in its intent when the wider public is either not interested or not skilled enough to perform this type of discussion.

Still, transparency is the most widely invoked principle in AI ethics guidelines: very few guidelines do not mention it.[92] And, it was mentioned in all of the guidelines reviewed specifically for this report.

Transparency about ADMS—or rather, the socio-technical system of which ADM is only a part—is invoked in relation to different elements:

- the existence of ADMS,[93] including their purpose, reach, and actual use[94] *(questions 2.1 and 2.2),*
- the definitions of key concepts and key measures (e.g., of 'automated decisions' or 'AI',[95] of fairness[96]) *(questions 2.8, 2.10, 2.12),*

Role of Transparency	Goals	Ethical Value
Transparency as disinfectant	Accountability, avoiding unethical behavior	Harm prevention
Transparency for crowd-sourcing	Collecting the opinions of experts and lay people, improving the technology	Beneficence
Transparency for informed choice	Enabling informed individual choice	Autonomy (individual)
Transparency for informed public debate	Enabling informed democratic deliberation	Autonomy (collective), democracy

Table 1: Different roles and goals assigned to transparency and the values that are to be realized through it.

91 Zarsky, "Transparent Predictions"; Felzmann et al., "Transparency You Can Trust".

92 Jobin, et al., "The Global Landscape of AI Ethics Guidelines".

93 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI"; Cities for Digital Rights, "Declaration of Cities Coalition for Digital Rights"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Identifiability of Algorithmic Decision-making"; Dataethical Thinkdotank, "White Paper," cf. "Fair Communication".

94 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI"; Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. "1. Zielorientierung", "8. Nachvollziehbarkeit".

95 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI".

96 Leslie, "Understanding Artificial Intelligence Ethics and Safety".

- the ethical or impact assessment concerning them,[97] their justification[98] *(questions 2.18 and 2.19),*
- the underlying data types and processing methods,[99] and
- their overall quality, characterized as accuracy,[100] effectiveness, efficiency,[101] or ability to support the administration[102] *(questions 2.16 and 2.17*).

In comparison to earlier guidelines examined in previous reviews,[103] the guidelines considered here seemed to put less emphasis on transparency in the sense of providing individual explanations of the causes or reasons why a specific decision was taken by an ADMS,[104] which is the subject of most discussions in the nascent field of [e]X[plainable]-AI.

While some guidelines acknowledge that an identical level of transparency is not always reasonable for all systems,[105] the level of transparency should be as high as possible within the constraints imposed by trade-offs with other goals.[106] The intended audience of transparency-related communication may vary: the individuals involved or affected,[107] the broader public,[108] or independent experts.[109] The format of the communication may also vary depending on the context—although this is rarely specified—for example, from a plain language notice on a website (e.g., declaring that an automated decision is made) to a fully documented report.[110] Even reports by whistle-blowers are sometimes considered deserving of protection and of encouragement by means of legal regulation and of the organizational structures of companies.[111]

Transparency can be considered the cornerstone of the approach developed here, which requires public

97 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Expertise and Oversight"; Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "Appendix C – Notice".

98 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI"; Government Digital Service and Office for Artificial Intelligence, UK, "A Guide to Using Artificial Intelligence in the Public Sector," cf. "Transparency"; Schweizerische Eidgenossenschaft – Der Bundesrat., "Leitlinien ‚Künstliche Intelligenz' für den Bund," cf. "Leitlinie 3".

99 Council of Europe, "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment"; Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "Appendix C – Notice".

100 Dataethical Thinkdotank, "White Paper," cf. "Fair communication".

101 Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "Reporting: 6.5.1".

102 Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "Appendix C – Notice".

103 Loi, "People Analytics Must Benefit the People"; Loi, et al., "A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data".

104 Government Digital Service and Office for Artificial Intelligence, UK, "A Guide to Using Artificial Intelligence in the Public Sector," cf. "Transparency"; Dataethical Thinkdotank, "White Paper," cf. "Transparency"; World Economic Forum, "AI Government Procurement Guidelines," cf. "Human in the Loop"; Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "6.2.3".

105 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Levels of Transparency"; Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," cf. "Transparency".

106 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Levels of Transparency".

107 Dawson, et al., "Artificial Intelligence: Australia's Ethics Framework"; Dataethical Thinkdotank, "White Paper," cf. "Transparency"; Government Digital Service and Office for Artificial Intelligence, UK, "A Guide to Using Artificial Intelligence in the Public Sector," cf. "Ongoing Review"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Expertise and Oversight"; Schweizerische Eidgenossenschaft – Der Bundesrat., "Leitlinien ‚Künstliche Intelligenz' für den Bund," cf. "Leitlinie 3".

108 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Public Debate"; Automated Decision Systems Task Force, "New York City Automated Decision Systems Task Force Report," cf. "Available Information".

109 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Expertise and Oversight"; AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI"; Schweizerische Eidgenossenschaft – Der Bundesrat., "Leitlinien ‚Künstliche Intelligenz' für den Bund," cf. "Leitlinie 3".

110 Government of Canada and Treasury Board Secretariat, "Directive on Automated Decision-Making," cf. "6.2", and "Appendix C – Notice".

111 Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Advancement of Public Benefit".

servants to prepare a transparency report as the main deliverable of all activities undertaken to ensure ethical standards. From this perspective, the desirable level of transparency is relative to the context and nature of the ADMS in question. However, rather than relying on a risk score, the present approach provides specific indications on the elements of transparency a public administration has to deal with or the issues and aspects it needs to make transparent.

/ Accountability

Accountability includes actions, choices, frameworks, and organizational structures designed to facilitate the identification and distribution of responsibility. Only human and legal persons can be held accountable, ADMS cannot. At a minimum, accountable agents are able to take responsibility for their actions and can be subjected to sanctions. Thus, promoting accountability is equivalent to promoting the capacity to identify who is responsible for what and who should be sanctioned (not only legally but also organizationally and through negative reputational consequences) for unethical or illegal outcomes.

The guidelines examined here require that accountability be promoted:

1. by assigning responsibility: it should be possible to identify who is responsible for ensuring that outcomes and behaviors comply with ethical standards (forward-looking responsibility),[112] and who is liable to sanctions if they fail to comply;[113]

2. through adequate structures and adequate organization of the processes of data science behind ADMS and of automation processes.[114] Organizations should "establish a continuous chain of responsibility for all roles involved in the design and implementation lifecycle of the project".[115] In this respect, it is important to emphasize that a continuous chain of responsibility for all roles requires clearly documented, monitored, and controlled processes and outcomes.[116] In other words, adequate structures for ensuring accountability are structures of control, as defined above (section 2.1). For the sake of brevity, the elements of control that have already been described in section 2.1 will not be repeated here.

3. by enabling that decisions taken by apparently impersonal systems can be contested or challenged[117] (in some cases by making use of the concept of due process),[118] or that the ADMS can be challenged in general, by virtue of its harmful or discriminatory effects (with concepts akin to public participation processes)[119]

4. by requiring that institutions deploying the ADMS are responsible for collecting the feedback of the people affected by it and for implementing the required remedial actions.[120]

---

112 Dawson, et al., "Artificial Intelligence: Australia's Ethics Framework," cf. "Accountability"; Schweizerische Eidgenossenschaft – Der Bundesrat., "Leitlinien ‚Künstliche Intelligenz' für den Bund," cf. "Leitlinie 4"; Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. „4. Projektmanagement".

113 Government of New Zealand, "Algorithm Charter for Aotearoa New Zealand," cf. "Human Oversight and Accountability".

114 Engelmann and Puntschuh, "KI im Behördeneinsatz," cf. „4. Projektmanagement".

115 Government Digital Service and Office for Artificial Intelligence, UK, "A Guide to Using Artificial Intelligence in the Public Sector," cf. "Accountability".

116 Government Digital Service and Office for Artificial Intelligence, UK, "A Guide to Using Artificial Intelligence in the Public Sector," cf. "Accountability".

117 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Contestability"; Dawson et al., "Artificial Intelligence: Australia's Ethics Framework," cf. "Contestability".

118 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI".

119 Cities for Digital Rights, "Declaration of Cities Coalition for Digital Rights".

120 AI Now Institute, et al., "Using Procurement Instruments to Ensure Trustworthy AI," cf. "Participatory Democracy, Diversity and Inclusion"; Council of Europe, "Recommendation CM/Rec(2020)1 on the Human Rights Impacts of Algorithmic Systems," cf. "Consultation and Adequate Oversight"; Dawson et al., "Artificial Intelligence: Australia's Ethics Framework," cf. "Recourse"; Automated Decision Systems Task Force, "New York City Automated Decision Systems Task Force Report," cf. "Impact Address".

5. by requiring that harm and damage due to unethical behavior be compensated.

The practical guidelines proposed below address the principle of accountability by turning it into an object of transparency. Checklist *items 2.3 to 2.6* require that individual responsibilities mentioned above be declared in the transparency report. Checklist *items 2.15* and *2.10* address the existence of structures by means of which decisions made by ADMS can be challenged.

3 Checklists

3.1 Introduction

The following two checklists provide a tool to generate transparency of technological automation projects and applications within public administration.

As to methodology, it consists in completing a written transparency report, demonstrating that the most important ethical issues have been both acknowledged and brought under human control, and that adequate accountability for the process has been ensured.

In order to examine the ethical requirements for the use of ADMS, two assessments should guide the practical activities of public administrations. Accordingly, two checklists are provided.

To some extent, the checklists can consider aspects that are already covered by other mechanisms a particular public administration might have put into operation (for example, rules and procedures in the fields of data protection or cybersecurity). This should certainly be taken into account when implementing the checklists in a specific context, but it cannot be considered at a general level here.

Based on the first checklist (triage checklist), the administration assesses which ethical transparency issues are worth documenting in detail during the project execution, and it adopts proportionate procedures that allow for generating the data and assessments needed to fill the report with informative content. This helps to determine:

a) how many ethical transparency items must be managed;

b) how many ethical transparency procedures need to be put in place;

c) how many resources need to be allocated for ethical transparency procedures;

d) which ethical transparency issues to address in detail in the report (and whether such a report is necessary at all).

The second checklist (transparency report checklist) serves as a guide to writing a highly detailed transparency report (in what follows "transparency report").

The transparency report can only be completed at the end of the development and implementation of an ADMS, hereafter referred to as the "project" (including the interaction of the socio-technical system with the target public in cases where the assessment of ethical issues requires such monitoring). However, the writing of the transparency report has to have already begun during the project: some of the information required for the transparency report can only be generated during the different phases of the project execution, and can no longer be generated when it has already been finalized. Thus, the transparency report checklist also includes indications as to the stages in the process at which specific information necessary for transparency must be generated. At the end of a project, the transparency report must include clear information about the processes that were put in place to address the specific ethical issues highlighted by checklist 1 (triage checklist).

If processes are altered after the system's implementation, the administration needs to review whether its

initial assessment is still valid or whether additional ethical transparency issues have emerged.

These ethical transparency processes should be activated by the administration and address all potential transparency issues indicated by the checklist items. If the administration is not able to provide an adequate degree of transparency about such ethical issues, or if the outcome of the transparency mechanisms demonstrate the inadequacy of the socio-technical system in a way that is incompatible with the reputation of the agency sponsoring it, this should result in a reevaluation of the project's goal or in investing more resources to find a feasible solution.

For present purposes, transparency is understood as communication to different audiences. The checklists themselves do not provide indications of a specific addressee for the transparency documentation. The reason for this is that, for transparency to be an element of accountability, these addressees must be those which, by virtue of their social role, are legally entitled to access this information, not merely permitted to do so.[121] The determination of these legal entitlements, from both the moral and legal point of view, is an exceedingly complex task, involving a sensible balance of disparate moral, organizational, and legal considerations.

At the same time, we have specific policy recommendations on this issue. In order to achieve meaningful transparency, all ADMS deployed in the public sector should be disclosed within a public register. These registers should contain information on the purpose, the underlying model, the actors involved in developing and deploying the system, and the results of the impact assessment—that is, in this case, the transparency report. Thus, for many systems, the audience will be the broad public (everyone).

Such registers also enable access for public interest research (by academics, civil society, and journalists) to relevant data on ADMS, allowing for an evidence-based debate on the automation of the public sector—which, in turn, serves as a prerequisite for exerting democratic control and ensuring accountability.

However, there can be legitimate reasons for omitting full public disclosure (for example, the protection of personal data). For such applications or categories of information, transparency must be provided toward a specific audience, e.g. the appropriate oversight institution, a superior within the same department, specific other public agencies, independent experts, committees, or stakeholder representatives to which the documentation will be communicated confidentially. In turn, these specific audiences should be publicly communicated within the register.

3.2 Checklist 1: Transparency Triage for ADMS

3.2.1 Introductory Remarks

This checklist should be answered as early as possible, i.e., during the planning stage, because it allows additional specifications for the project (besides its primary goal) to be considered. Checklist 1 helps to determine the main transparency issues that have to be documented. It should not be considered an exhaustive list of all potential risks for all contexts (other risks to people, assets, and societies as a whole, not mentioned here, may exist). It is a heuristic methodological tool, not a guarantee that all problems will be considered. Checklist 1 does not specify who should answer the checklist and who should own the processes and decisions that checklist 2 mentions. This is due to the task of assigning legal entitlements to access the information mentioned above.[122]

Checklist 1 is not restricted to specific technologies or ADMS. On the contrary, it should help in identifying the kinds of ADMS that display an increased need

121 Michele Loi and Matthias Spielkamp, "Towards Accountability in the Use of Artificial Intelligence for Public Administrations," https://arxiv.org/abs/2105.01434, 2021, https://doi.org/10.1145/3461702.3462631.

122 Section 3.1, last paragraph.

for transparency. It is easier to achieve the morally required level of transparency by means of the checklist if:

- the project takes ethical issues into consideration in early planning and implementation of the product;
- specifications intended to mitigate such ethical issues are included from the beginning (ethics by design approach);
- information about the measures taken to deal with ethical issues is duly documented at every stage in the process;
- transparency and accountability of the procedures run by people exist in all phases of the project.

The checklist implies two levels of transparency: low detail transparency and high detail transparency:

- Low detail transparency: document and store only the answers to the triage checklist, including the reasons in support of your answers.
- High detail transparency: file a detailed transparency report (checklist 2). The answers to the questions in the triage checklist determine whether a transparency report must be filled, and which sections of the transparency report must be filled.

Stage of assessment for triage checklist 1: at the very beginning of an automation project.

/ Harm Prevention[123]

1.1 *Enhanced privacy harm:* Does the decision deal with special categories of personal data, as defined by applicable legal norms?

1.2 *Cybersecurity harm:* May malicious parties have especially strong motives to hack the system? Can they easily achieve substantial financial gain—including by means of blackmailing—or can a hacked system be used to achieve political goals (including expressing political opposition against the system)?

1.3 Is the socio-technical system used to take, recommend, or affect decisions about individuals in a way that influences the outcome, i.e., what decision is taken?
[For example, an automated spell checker is part of a socio-technical system. If it is used by people who take decisions about individuals, it can be described as "used to take decisions about individuals". At the same time, such a tool does not influence (in any detectable and scientifically plausible way) what decision is taken.]

1.4 *Fundamental right harm:* Is the system used to take a decision about a legal duty or right of an individual? [For the meaning of "used to take a decision" in this context, see explanation provided for *question 1.3*]

1.5 *Fundamental right erosion harm:* Does the system make it more or less probable that certain individuals will enjoy the substance of a right? Or does the system make it more or less probable that certain individuals will be sanctioned? Or does the system affect the probability that an individual case will attract the attention of, or be ignored by, the administration?

1.6 *Avoidability:* Can individuals avoid the decision or demand that the decision is taken via a different procedure, not involving the same technical system?

1.7 *Identifiability:* Can the person about whom a decision is taken with the help of the relevant

123 The checklist does not include the value of beneficence because it is not included among the values of the Independent High-Level Expert Group On Artificial Intelligence Set Up By The European Commission, "Ethics Guidelines for Trustworthy AI", which we follow here (see section 2.1).

tool prove the wrongness of the decision without going to court?

1.8 *Reversibility:* Is the harm of a wrong decision fully reversible?

1.9 *Compensability:* Is it possible to compensate the individual or family fully and adequately for the harm of a wrong decision, when it is ascertained that the decision was wrong and cannot be reversed?

1.10 *Public good harm:* Does the decision concern any of the following areas of public life or public sector resources:
- the administration of justice,
- access to educational opportunities,
- access to democratic processes,
- access to social welfare services, including healthcare,
- public health measures,
- the environment?

1.11 When the ADMS used by the public administration is delegated to a private third-party (in its entirety or for aspects of its implementation), does this result in a change in:
- public computing infrastructure,
- public data assets, or
- intangible assets (e.g., competences) in the public sector?

/ Justice and Fairness

1.12 *Political risk:* Is it possible that the technical system will have an effect on a political decision (e.g., a popular vote)?

1.13 *Economic risk:* Does the technical system affect the distribution of public resources to economic actors in society?

1.14 *Statistical proxy risk:* Does the technical system rely on a statistical model of human behavior or personal characteristics?

1.15 *Procedural regularity risk:* Is the system designed to be adaptive so that it will not treat all new cases in the same way as those it encountered in the past, because it changes its parameters (e.g., in order to become more efficient)?

/ Autonomy

1.16 *Human replacement:* Is the goal of the technical tool to automate a fully deterministic system of rules, which requires minimal creativity and human judgment by current human operators and which does not involve estimating risk or probability?

1.17 *Complexity risk:* Does the technical system rely on parameters, features, factors, or decision criteria that do not correspond to those normally considered by most experts in the field?

1.18 *Ignorance risk:* Does the technical system make judgments (predictions or recommendations) that the public administration employees in charge lack the competence (as opposed to the power) to criticize and overrule?

1.19 Does the technical system rely on third-party infrastructure the public entity has no unrestricted control over and/or access to, e.g., data sets or computing power?

3.3 Checklist 2: Transparency Report

The goal of the second checklist lies in achieving transparency in order to foster the trustworthiness of the process. This goal is achieved by means of completing a report.

Instruction for use: If answering checklist 1 indicates that a transparency report must be completed, the flow chart determines which set of questions in checklist 2 the transparency report must address.

/ Section 1

Stage of assessment for checklist items 2.1 to 2.6: before you design your system

/ Value Transparency

2.1 What is the problem the system is designed to solve?

2.2 What are further desiderata of the system? Consider at least:

2.2.1 *Privacy*
[In this part of the report, please specifically address the issues mentioned in checklist 1—*question 1.1*]

2.2.2 *Cybersecurity*
[In this part of the report, please specifically address the issues mentioned in checklist 1—*question 1.2*]

2.2.3 *Fairness*
[In this part of the report, please specifically address the issues mentioned in checklist 1—*questions 1.12 to 1.15*]

2.2.4 *Explainability*
[In this part of the report, please specifically address the issues mentioned in checklist 1—*questions 1.17 and 1.18]*

/ Accountability Transparency

[We recommend that this section is filled in every case. If you answer "yes" to *question 1.16,* you must show that the new responsibilities provide at least equivalent possibilities for challenging decisions as the system previously in place.]

2.3 Who is responsible for the design of the system (level of project organization)?

2.4 Who is responsible for the implementation of the system and for its outcomes (level of home organization)?

2.5 Who is responsible for managing responses and feedback by end-users, i.e., the people using or being assisted by the system?

2.6 Who is responsible for answering to doubts or challenges by individuals affected by the use of the system?

/ Section 2

Stage of assessment for checklist items 2.7 to 2.19: after testing the system

/ Translation and Control Transparency

2.7 What methodologies have been used to test and measure the performance of the system?
[Please indicate how you measure the performance with respect to the main goal of the system, specified in checklist 2—*question 2.1*]

2.8 What methodologies have been used to identify …

2.8.1 the stakeholders directly affected by the system's predictions/recommendation/decisions? What are the foreseen effects on these individuals?

2.8.2 the individuals affected by digital transformation in the public administration (e.g., public administration personnel)? What are the foreseen effects on these individuals?

2.9 What protocols are in place to deal with system errors and malfunction?

2.10 What methodologies have been used to define and protect privacy?
[In this part of the report, please specifically address the risks mentioned in checklist 1—*question 1.1*]

2.11 What cybersecurity measures are in place?
[In this part of the report, please specifically address the issues mentioned in checklist 1—*question 1.2*]

2.12 What methodologies have been used to define and measure the bias and fairness of the system?

[In this part of the report, please specifically address the issues mentioned in checklist 1—*questions 1.12 to 1.15*. The issue mentioned in *questions 1.14 and 1.15* must always be explicitly addressed in this section of the report, even when the writing of section 2.12 is not triggered by answering "yes" to those questions.]

2.13 How are individual predictions/recommendations/decisions of the system explained to system end-users and individuals affected by the use of the system?

[In this part of the report, please specifically address the issues mentioned in checklist 1—*questions 1.17 and 1.18*]

2.14 Is system deployment continuously monitored after the testing phase …

a) at all times?
b) within a given timeframe?
c) through which measures?

2.15 Are there options for people affected by a decision to learn about the output of the automated system and to challenge predictions/recommendations/decisions influenced by the system?

[If pertinent, describe how these options are related to organizational roles mentioned in checklist 2—*questions 2.5 and 2.6*]

/ Performance Transparency

According to the tests run so far:

2.16 How does the system perform relative to the chosen relevant metrics?

[Please refer to all the goals and desiderata described when answering checklist 2—*questions 2.1 and 2.2*]

2.17 How does the system compare to the one previously in place, if any, or to established benchmarks, if available?

[Please also refer to the goals and desiderata specified in checklist 2—*questions 2.1 and 2.2*]

2.18 What are the remaining security and privacy risks and why are they reasonable?

[In this part of the report, please address all the desiderata identified in checklist 2—*questions 2.1 and 2.2* as well as *questions 2.10 and 2.11*]

2.19 Please describe relevant unresolved biases or possible sources of unfairness in the system and explain why they cannot be resolved (e.g., by explaining trade-offs with other system goals, including conflicting fairness goals).

[In this part of the report, please address all the desiderata identified in checklist 2—*question 2.2.3* and explain how the issues identified in checklist 1—*questions 1.12 to 1.15* have been addressed.]

/ Section 3

Stage of assessment for checklist item 2.20: after system implementation, when the system is monitored

Transparency about the solutions offered for addressing ethical issues may sometimes require the continuous monitoring of the project, which also goes beyond the testing phase. Such continuous monitoring implies that the transparency report has to be updated.

2.20 During monitoring, have predictions/recommendations/decisions by the system ever been challenged …

a) by system end-users?
b) by individuals subjected to decisions?

4 Example of the Use of Checklists 1 and 2: Swiss COMPAS

In what follows, the application of the triage for ADMS (checklist 1) and of the requirements for the transparency report (checklist 2) will be illustrated by way of a fictional example: the Swiss COMPAS system, a recidivism prediction tool (similar to the US-American system COMPAS).[124] This imaginary tool would be a tool used by a Canton in Switzerland to determine, based on data about a convicted criminal, the probability that this person will reoffend once released from prison.

4.1 Introductory Remarks

This hypothetical transparency report is only provided as an illustration of what a real, fact-based report may look like. The case at hand is not meant to provide a model of "ethical best practice": transparency is achieved when the situation concerning automation is portrayed honestly, including aspects that may not have been dealt with in an ethically adequate way. We assume that most real-world cases will be imperfect from the point of view of ethical adequacy: transparency is meant to enable gradual, incremental improvements.

Moreover, this hypothetical report is conceived as an imaginary "first draft" produced by the Canton, which can and would be improved by interactions with stakeholders demanding more information than the current report gives in subsequent interactions. Transparency is not meant to provide ultimate moral justifications for the use of such technologies but, instead, to make them gradually improvable.

The flowchart will lead users to fill the transparency report in a different order than the order in which the different sections of the transparency report will be read in the final report.

The structure of the transparency report is always the same, but the triage question and the flowchart commands determine which sections must be filled in a specific case, and which can be left out. This may differ from case to case, depending on the answers to the triage questions.

The flowchart may indicate that certain questions must be answered more than once (e.g., in response to different triage questions). Such repetitions can be ignored. In this example, we do not mention any repetition produced by flowchart requests.

The section "Checklists 1 and 2" below illustrates how a person or a team in charge of developing a transparency report may answer the questions based on the checklists. The section "Transparency Report" shows the report in a version ready for publication.

124 https://en.wikipedia.org/wiki/COMPAS_(software).

4.2 Checklists 1 and 2

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
1.1. Does the decision deal with special categories of data, as defined by applicable legal norms?	The Swiss COMPAS assessment tool requires data on the gender of the person assessed. This declaration is protected by anti-discrimination laws and, in certain cases, prohibited. The tool also collects confidential information, such as whether the parents of the person being assessed are separated, or if friends and relatives have criminal records.	**2.2.1. What are the further desiderata of the system with respect to privacy?** [Declare privacy desiderata of the system. What privacy protections ought to be in place? Consult e.g., with the data protection department to write this part of the report.] **2.8.1 What methodologies have been used to identify the stakeholders directly affected by the system's predictions/recommendation/decisions? What are the foreseen effects on these individuals?** Example: "We ran a brainstorming meeting with prosecutors and judges of the Canton and criminal justice lawyers. During this meeting, the stakeholders directly affected by predictions were identified: the defendants, their defense lawyers, their families, potential victims (if defendants reoffend), and communities in which people who may reoffend end up living. Our analysis of stakeholder interests is as follows: A) Defendants. The tool is in the interest of defendants who are unlikely to reoffend (or who are statistically indistinguishable from those unlikely to reoffend). It is especially in the interest of those least able to exercise their right to obtain a favorable parole verdict in that they cannot afford the best lawyers. It is not in the interest of those with good lawyers who can get more favorable parole conditions for them. B) Defendant's lawyers. The tool is against their interest as it will be an element in the judge's decision that lawyers are not able to challenge. C) Families. Families of defendants will benefit from the higher probability of the defendant being released on parole - if the use of Swiss COMPAS leads to an increase in parole approvals compared to the status quo (unless the defendant is charged for a crime against his or her family). This is strictly related to the proportion of decisions to approve parole. ▶

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
		D) Potential victims. They will benefit from Swiss COMPAS if it leads to a smaller proportion of reoffending defendants being released. Their interests are not necessarily promoted if Swiss COMPAS leads to a lower proportion of parole approvals. On the one hand, if the tool is less accurate than human judges, a lower proportion of parole approvals may cause the rate of reoffending parolees to grow. On the other hand, if the tool is more accurate than human judges, a higher proportion of parole approvals may go hand in hand with a reduction of crime from reoffending parolees. E) The interest of the communities that may benefit can be considered as the combination of the interests of the family members of the defendants (as specified in C above) and of individuals whose interests are aligned to those of the family members, as well as the interests of the potential victims of parolees (as specified in D above) and of individuals whose interests are aligned with those of these potential victims (e.g., the victim's children). A tool that is able to release more people on parole while simultaneously reducing the incidence of reoffending parolees should be welcomed by the communities of the defendants." **2.10. What methodologies have been used to define and protect privacy?** [Here you explain methodologies actually used to protect personal data with respect to privacy.] **2.18. What are the remaining security and privacy risks and why are they reasonable?** [Here you explain why the cybersecurity risk resulting from *2.9* and *2.11* is considered reasonable given the stakes and probabilities of an attack.]

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
1.2 May malicious parties have especially strong motives to hack the system? Can they easily achieve substantial financial gain—including by means of blackmailing—or can a hacked system be used to achieve political goals (including expressing political opposition against the system)?	Because the data collected is sensitive, cybersecurity must be high to prevent attacks by motivated intruders.	**2.2.2 What are the further desiderata of the system with respect to cybersecurity?** [Declare cybersecurity desiderata of the system. Request technical expertise from cybersecurity experts.] **2.9 What protocols are in place to deal with system errors and malfunction?** Example: "Wrong predictions about persons released from prison—who were considered low-risk but reoffend—are recorded through a specific penal law procedure that monitors parolee's success. All data is securely stored in X [specify databases or registries] and accessible to Y [specify roles of the public administration]. The measure taken to address false prediction is: penal law already foresees legal consequences to reoffending while on parole. The proportion of defendants for which parole was approved and rejected is recorded and made accessible to X, Y [roles in the organization]. The possibility of false predictions (in the sense that their risk score is based on data correction error) has been considered and the following plan has been developed in order to minimize this type of error [specify procedure to minimize error, if any]." **2.11 What cybersecurity measures are in place?** [In this section, you explain cybersecurity safeguards built into the system. Optimally, this section is drafted by cybersecurity experts.] **2.18 What are the remaining security and privacy risks and why are they reasonable?** Answer: see above *(question 1.1, item 2.18)*
1.3 Is the socio-technical system used to take, recommend, or affect decisions about individuals in a way that influences the outcome, i.e., what decision is taken?	Yes, the system produces scores that are used by judges to decide whether to grant parole to the defendant.	**2.2.3 What are the further desiderata of the system with respect to fairness?** [This aspect is too complex to fully present here. It requires a joint analysis with domain experts (at least statistics experts, criminal justice scholars, penal law lawyers, social justice experts) and affected stakeholder representatives able to present and defend a reasoned view of what they take to be fair and unbiased judgment.] Example: "We define fairness as follows: On average, recidivism predictions are equally precise for men and women. The justification for defining fairness this way is: ..." ▶

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
		2.2.4 What are the further desiderata of the system with respect to explainability? As an example of the considerations relevant for this section of the report, consider the following: "Since this is a high-stake decision ultimately taken by a human being (the judge), it seems important that this human being can form a mental model of the factors considered and their respective weight in the generation of the outcome (i.e., the score). If the Swiss COMPAS tool uses a secret algorithm, this requirement may not be met. On the other hand, if the formula were publicly known, parolees would be inclined to answer insincerely to improve their risk scores. (The US COMPAS algorithm is secret.)" Ideally, a genuine report would involve a more thorough analysis and discussion of this trade-off. It would also include potentially feasible solutions how to avoid it. **2.7 What methodologies have been used to test and measure the performance of the system?** [Please indicate how you measure the performance with respect to the main goal of the system, specified in check-list 2—*question 2.1*] Example: "We tested the performance of the algorithm by checking its predictions on the basis of historical data on arrests of parolees, which the cantonal police collected. We measured the overall accuracy of the predictions, the false negative rate and the false positive rate. The technical attachment is available [here]". **2.8.1 What methodologies have been used to identify the stakeholders directly affected by the system's predictions/recommendation/decisions? What are the foreseen effects on these stakeholders?** Answer: see above *(question 1.1, item 2.8.1)* **2.9 What protocols are in place to deal with system errors and malfunction?** Answer: see above *(question 1.2, item 2.9)*
1.4 Is the system used to take a decision about a legal duty or right of an individual?	Yes, the system is used to take a decision about being allowed to leave prison.	

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
1.6 Can individuals avoid the decision, or demand that the decision be taken via a different procedure, not involving the same technical system?	Two scenarios: A. Yes, defendants can avoid being assessed by Swiss COMPAS by requesting this through their lawyer, and this will always be accepted. As a consequence, his or her recidivism probability will be assessed by a human judge without the help of the tool. B. No, every defendant is required to accept the assessment.	Scenario A) continue with *question 1.10.* Scenario B) continue with *question 1.7.*
1.7 Can the person about whom a decision is taken with the help of the relevant tool prove the wrongness of the decision without going to court?	No, a decision to deny parole cannot be shown to be based on a wrong prediction. Showing that someone, who was previously released on parole, has reoffended does not show that the decision was wrong, because it was based explicitly on an uncertain risk evaluation, which is (sometimes) compatible with a low-risk defendant reoffending.	**2.5 Who is responsible for managing responses and feedback by end-users, i.e., the people using or being assisted by the automation?** Example: "Judges can request an informative exchange with an expert from the Swiss COMPAS company, who can explain in lay terms what the tool considers and why, or which kind of bias the system involves." **2.6 Who is responsible for answering to doubts or challenges by individuals affected by the use of the system?** Example: "No one." **2.14 Is system deployment continuously monitored after the testing phase …** **a) at all times?** **b) within a given timeframe?** **c) through which measures?** Example: "No. No monitoring system is in place to track the performance of the system over time." **2.15 Are there options for people affected by a decision to learn about the output of the automated system and to challenge predictions/recommendations/ decisions influenced by the system?** Example: "No, no such system has been designed and put into operation." ▶

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
		2.17 How does Swiss COMPAS compare to the system previously in place, if any, or to established bench-marks, if available? Example: "The system previously in place was a parole decision by human judges, taking into consideration X, Y, Z. Unfortunately, there are no reliable studies about the overall accuracy of parole decisions by Swiss human judges (i.e., about the proportion of people released on parole who reoffend). There are some studies in the US showing that predictive tools used in the US are x% more accurate than human judges, but it is unclear whether and how this transfers to the Swiss context, since the performance of Swiss judges is not known. To measure the accuracy of Swiss COMPAS, we assume that a recom-mendation to deny parole is taken for any defendant with a Swiss COMPAS risk score higher than 0.3 regardless of gender (i.e., if this threshold is used, out of 1000 people released on parole, approximately 300 would reoffend). With this threshold, the Swiss COMPAS tool has an accuracy of x%, which is lower than the accuracy of the US COMPAS tool, which is y%. However, the false positive rate (the proportion of people predicted to reoffend in the test data who have not actually reoffended) is much lower in Swiss COMPAS than US COMPAS, while the false negative rate is slightly higher. Hence, compared to its US counterpart, Swiss COMPAS performs better in identifying people who would not reoffend and deserve to obtain parole, but it performs worse in identifying people who reoffend and should be kept in prison." **2.20 During monitoring, have predictions/ recommendations/decisions by the system ever been challenged …** **a) by system end-users?** **b) by individuals subjected to decisions?** Example: "The system is not yet in operation in Switzer-land. It is possible that similar systems may have been challenged in other countries in which they are deployed, but our department does not possess this information."

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
1.10 Does the decision concern any of the following areas of public life or public sector resources: **the administration of justice,** **access to educational opportunities,** **access to democratic processes,** **etc.?**	Yes, it is used for administering justice.	**2.8.1 What methodologies have been used to identify the stakeholders directly affected by the system's predictions/recommendations/decisions? What are the foreseen effects on these stakeholders?** Answer: see above *(question 1.1, item 2.8.1)* **2.20 During monitoring, have predictions/ recommendations/decisions by the system ever been challenged …** **a) by system end-users?** **b) by individuals subjected to decisions?** Answer: see above *(question 1.7, item 2.20)*
1.11 Does the acquisition or the deployment of the ADMS result in a change in: **– public computing infrastructure,** **– public data assets, or** **– intangible assets (e.g., competences) in the public sector?**	Yes.	**2.2.3 What are the further desiderata of the system with respect to fairness?** Answer: see above *(question 1.3, item 2.2.3)* **2.8.2 What methodologies have been used to identify the individuals affected by digital transformation in the public administration (e.g., public administration personnel)? What are the foreseen effects on these individuals?** Example: "A meeting of heads of departments X, Y, Z was organized. The following stakeholders and respective effects on them have been identified: 1) Judges: the use of the Swiss COMPAS algorithm is highly welcomed by most parole judges for two reasons. First, they feel they have insufficient time to review the profiles of defendants when deciding on parole, given the high number of cases that need to be reviewed and the pressure to devote more time on more consequential decisions such as the sentence for a crime. Second, partly as a result of the short time they are expected to devote to this task, they complain about the subjectivity and poor accuracy of their estimations. Accordingly, they hope that the accuracy of their predictions can be both improved and made more objective, or, at least, uniform. Based on an internal poll, only a minority of these judges cares about the logic behind the risk assessment being made accessible to them. The majority does not consider this essential as long as there is strong evidence that the tool is accurate. ▶

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
		2) Duty cantonal lawyers: the duty cantonal (defense) lawyers do not welcome this move. They complain about the opaqueness of the tool, which, they claim, makes it impossible for them to defend the rights of the people they assist. They plan to take legal action if the algorithm behind the score is not made accessible—and they are ready to appeal against corresponding decisions up to the highest court of justice."
1.12 Is it possible that the technical system will have an effect on a political decision (e.g., a popular vote)?	No.	
1.13 Does the technical system affect the distribution of public resources to economic actors in society?	No.	
1.14 Does the technical system rely on a statistical model of human behavior or personal characteristics?	Yes.	**2.2.3 What are the further desiderata of the system with respect to fairness?** Answer: see above *(question 1.3, item 2.2.3)* **2.8.1 What methodologies have been used to identify the stakeholders directly affected by the system's predictions/recommendations/decisions? What are the foreseen effects on these stakeholders?** Answer: see above *(question 1.1, item 2.8.1)* **2.12 What methodologies have been used to define and measure the bias and fairness of the system?** Example: "We have measured the false discovery rate, false omission rate, false positive rate and false negative rate in the aggregate as well as disaggregated by gender, assuming judges will deny parole to any defendant with a risk score higher than 0.3. However, the false omission and false discovery rates are equal for both genders. Risk scores have the same predictive value (they offer equally strong evidence that a parolee will reoffend) irrespective of the gender for which they are used. The system is not designed to be adaptive, because an adaptive system would not have a homogeneous output for individuals with the same features. That is, people with the same characteristics may receive different recommendations, in a way that would be difficult to control by the judges." ▶

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
		2.13 How are individual predictions/recommendations/decisions of the system explained to system end-users and individuals affected by the use of the system? Example: "The judges have been informed of the features that are considered by the prediction tool. However, the formula and the respective weight of these features are a trade secret and for this reason, they have not been revealed to them." **2.19 Please describe relevant unresolved biases or possible sources of unfairness in the system and explain why they cannot be resolved (e.g., by explaining trade-offs with other system goals, including conflicting fairness goals).** Example: "It is impossible to equalize all above-mentioned rates for all groups, because average rates of recidivism differ for men and women. For this threshold value, the false positive rate and the false negative rate differ with respect to different genders. Female prisoners have a higher false negative rate but a lower false positive rate, i.e., they are less likely to be wrongly detained and more likely to be wrongly released compared to men. Male prisoners have a higher false positive rate and a lower false negative rate, i.e., they are more likely to be wrongly kept in prison and less likely to be wrongly detained. Given the fairness goals specified in 2.2.3 and the importance of equalizing the false discovery and false omission rate (as well as communicating calibrated scores to judges), which follows from this analysis of the goals of the system, we contend that this solution scores best from the point of view of fairness." **2.20 During monitoring, have predictions/ recommendations/decisions by the system ever been challenged …** **a) by system end-users?** **b) by individuals subjected to decisions?** Answer: see above *(question 1.7, item 2.20)*

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
1.15 Is the system designed to be adaptive so that it will not treat all new cases in the same way as those it encountered in the past, because it changes its parameters (e.g., in order to become more efficient)?	No.	
1.16 Is the goal of the technical tool to automate a fully deterministic system of rules, which requires minimal creativity and human judgment by current human operators and which does not involve estimating risk or probability?	No.	
1.17 Does the technical system rely on parameters, features, factors, or decision criteria that do not correspond to those normally considered by most experts in the field?	Assuming that "domain expert" = "judges", the answer is affirmative, because the system relies on features and factors that are usually not considered (or at least not systematically) by human judges.	**2.2.4 What are the further desiderata of the system with respect to explainability?** Answer: see above *(question 1.3, item 2.2.4)* **2.8.1 What methodologies have been used to identify the stakeholders directly affected by the system's predictions/recommendations/decisions? What are the foreseen effects on these stakeholders?** Answer: see above *(question 1.1, item 2.8.1)* **2.13 How are individual predictions/recommendations/decisions of the system explained to system end-users and individuals affected by the use of the system?** Answer: see above *(question 1.14, item 2.13)* **2.14 Is system deployment continuously monitored after the testing phase …** **a) at all times?** **b) within a given timeframe?** **c) through which measures?** Answer: see above *(question 1.7, item 2.14)* **2.20 During monitoring, have predictions/recommendations/decisions by the system ever been challenged …** **a) by system end-users?** **b) by individuals subjected to decisions?** Answer: see above *(question 1.7, item 2.20)*

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
1.19 Does the technical system rely on third-party infrastructure the public entity has no unrestricted control over and/or access to, e.g., data sets or computing power?	Yes, the algorithm producing the risk score is a trade secret owned by the Swiss COMPAS company.	**2.8.2 What methodologies have been used to identify the individuals affected by digital transformation in the public administration (e.g., public administration personnel)? What are the foreseen effects on these individuals?** Answer: see above *(question 1.11, item 2.8.2)*
Check: Did the chart conclude you should write a transparency report?	Yes.	**2.1 What is the problem the system is designed to solve?** Example: "Currently, there are doubts about the quality of parole decisions by judges, inter alia related to a disproportion between the lack of time available to judges to make these decisions and the number of cases they are required to adjudicate on. The intention of Swiss COMPAS is to provide a risk indication to the judge, combined with an optimized recommendation (to give or deny parole), which the judge is free to ignore. The goal of the system is to be both more accurate and more homogeneous in forecasting recidivism—thus less prone to subjective decisions—than the current assessment by human judges." **2.3 Who is responsible for the design of the system?** Example: "Gianni Contabene, CEO of the "Southpointe" company, with headquarters in Lugano, Switzerland." **2.4 Who is responsible for the implementation of the system and for its outcomes?** [Indicate the roles responsible for the parole procedures, e.g., in the Cantonal criminal law tribunal] **2.5 Who is responsible for managing responses and feedback by end-users, i.e., the people using or being assisted by the system?** Answer: see above *(question 1.7, item 2.5)* **2.6 Who is responsible for answering to doubts or challenges by individuals affected by the use of the system?** Answer: see above *(question 1.7, item 2.6)* **2.14 Is system deployment continuously monitored after the testing phase …** **a) at all times?** **b) within a given timeframe?** **c) through which measures?** Answer: see above *(question 1.7, item 2.14)* ▶

Triage Checklist Questions (checklist 1)	*Answer to Question*	*Implications for Transparency Report (based on questions in checklist 2)*
		2.15 Are there options for people affected by a decision to learn about the output of the automated system and to challenge predictions/ recommendations/decisions influenced by the system? Answer: see above *(question 1.7, item 2.15)* **2.16 How does the system perform relative to the chosen relevant metrics?** [Please refer to all the goals and desiderata described when answering checklist 2—*questions 2.1 and 2.2*] "The primary goal is to make predictions as accurate as possible: The system should minimize the number of people who are denied parole even though they would not reoffend as well as the number of those who are released on parole but who would reoffend. The system has an overall accuracy rate of x%. Its comparison with relevant national and international benchmarks is described in section 2.17 above." **2.17 How does the system compare to the one in place previously, if any, or to established benchmarks, if available?** Answer: see above *(question 1.7, item 2.17)* **2.19 Please describe relevant unresolved biases or possible sources of unfairness in the system and explain why they cannot be resolved (e.g., by explaining trade-offs with other system goals, including conflicting fairness goals).** Answer: see above *(question 1.14, item 2.19)* **2.20 During monitoring, have predictions/recommendations/decisions by the system ever been challenged …** **a) by system end-users?** **b) by individuals subjected to decisions?** Answer: see above *(question 1.7, item 2.20)*
Final request: Are there additional ethical issues? Please include your answer when completing section 2.2 and all subsequent questions referring to the ethical issues mentioned in 2.2.		Example: "Additional ethical questions are not known to us."

4.3 Transparency Report

The following section presents the comprehensive transparency report, which results from answering checklist 1 for the hypothetical example (Swiss COMPAS) outlined above. This specific example report contains answers to all questions, which, however, only serves an illustrative purpose. Usually and in real life scenarios, this is not expected: transparency reports should list all those answers that are triggered by the questions in checklist 1.

a) What is the system intended to deliver and which requirements with respect to fundamental rights protection should it meet?

2.1 What is the problem the system is designed to solve?

Example: "Currently, there are doubts about the quality of parole decisions by judges, inter alia related to a disproportion between the lack of time available to judges to make these decisions and the number of cases they are required to adjudicate on. The intention of Swiss COMPAS is to provide a risk indication to the judge, combined with an optimized recommendation (to give or deny parole), which the judge is free to ignore. The goal of the system is to be both more accurate and more homogeneous in forecasting recidivism—thus less prone to subjective decisions—than the current assessment by human judges."

2.2 What are further desiderata of the system?

2.2.1 What are further desiderata of the system with respect to privacy?

[Declare privacy desiderata of the system. What privacy protections ought to be in place? Consult e.g., with the data protection department to write this part of the report.]

2.2.2 What are further desiderata of the system with respect to cybersecurity?

[Declare cybersecurity desiderata of the system. Request technical expertise from cybersecurity experts.]

2.2.3 What are further desiderata of the system with respect to fairness?

[This aspect is too complex to fully present here. It requires a joint analysis with domain experts (at least statistics experts, criminal justice scholars, penal law lawyers, social justice experts) and affected stakeholder representatives able to present and defend a reasoned view of what they take to be fair and unbiased judgment.]

Example: "We define fairness as follows: On average, recidivism predictions are equally precise for men and women. The justification for defining fairness this way is: ..."

2.2.4 What are further desiderata of the system with respect to explainability?

As an example of the considerations relevant for this section of the report, consider the following: "Since this is a high-stake decision ultimately taken by a human being (the judge), it seems important that this human being can form a mental model of the factors considered and their respective weight in the generation of the outcome (i.e., the score). If the Swiss COMPAS tool uses a secret algorithm, this requirement may not be met. On the other hand, if the formula were publicly known, parolees would be inclined to answer insincerely to improve their risk score. (The US COMPAS algorithm is secret.)"
Ideally, a genuine report would involve a more thorough analysis and discussion of this trade-off. It would also include potentially feasible solutions how to avoid it.

b) Who is accountable?

2.3 Who is responsible for the design of the system (level of project organization)?

Example: "Gianni Contabene, CEO of the "Southpointe" company, with headquarters in Lugano, Switzerland."

2.4 Who is responsible for the implementation of the system and for its outcomes (level of home organization)?

[Indicate the roles responsible for the parole procedures, e.g., in the Cantonal criminal law tribunal]

2.5 Who is responsible for managing responses and feedback by end-users, i.e. the people using or being assisted by the system?

Example: "Judges can request an informative exchange with an expert from the Swiss COMPAS company, who can explain in lay terms what the tool considers and why, or which kind of bias the system involves."

2.6 Who is responsible for answering to doubts or challenges by individuals affected by the use of the system?

Example: "No one."

c) Transparency information on implementation and control of the system

2.7 What methodologies have been used to test and measure the performance of the system?

[Please indicate how you measure the performance with respect to the main goal of the system, specified in checklist 2—*question 2.1*]

2.8 What methodologies have been used?

2.8.1 What methodologies have been used to identify the stakeholders directly affected by the system's predictions/recommendations/decisions? What are the foreseen effects on these individuals?

Example:
"We ran a brainstorming meeting with prosecutors and judges of the Canton and criminal justice lawyers. During this meeting, the stakeholders directly affected by predictions were identified: the defendants, their defense lawyers, their families, potential victims (if defendants reoffend), and communities in which people who may reoffend end up living.

Our analysis of stakeholder interests is as follows:

A) Defendants. The tool is in the interest of defendants who are actually unlikely to reoffend (or who are statistically indistinguishable from those unlikely to reoffend). It is especially in the interest of those least able to exercise their right to obtain a favorable parole verdict in that they cannot afford the best lawyers. It is not in the interest of those with good lawyers who can get more favorable parole conditions for them.

B) Defendant's lawyers. The tool is against their interest as it will be an element in the judge's decision that lawyers are not able to challenge.

C) Families. Families of defendants will benefit from the higher probability of the defendant being released on parole, if the use of Swiss COMPAS leads to an increase in parole approvals compared to the status quo (unless the defendant is charged for a crime against his or her family). This is strictly related to the proportion of decisions to approve parole.

D) Potential victims. They will benefit from Swiss COMPAS if it leads to a smaller proportion of reoffending defendants being released. Their interests are not necessarily promoted if Swiss COMPAS leads to a lower proportion of parole approvals. On the one hand, if the tool is less accurate than human judges, a lower proportion of parole approvals may cause the rate of reoffending parolees to grow. On the other hand, if the tool is more accurate than human judges, a higher proportion of parole approvals may go hand in hand with a reduction of crime from reoffending parolees.

E) The interest of the communities that may benefit can be considered as the combination of the

interests of the family members of the defendants (as specified in C above) and of individuals whose interests are aligned to those of the family members, as well as the interests of the potential victims of parolees (as specified in D above) and of individuals whose interests are aligned with those of these potential victims (e.g., the victim's children). A tool that is able to release more people on parole while simultaneously reducing the incidence of reoffending parolees should be welcomed by the communities of the defendants."

2.8.2 What methodologies have been used to identify the individuals affected by digital transformation in the public administration (e.g., public administration personnel)? What are the foreseen effects on these individuals?

Example: "A meeting of heads of departments X, Y, Z was organized. The following stakeholders and respective effects on them have been identified:

1) Judges: the use of the Swiss COMPAS algorithm is highly welcomed by most parole judges for two reasons. First, they feel they have insufficient time to review the profiles of defendants when deciding on parole, given the high number of cases that need to be reviewed and the pressure to devote more time on more consequential decisions such as the sentence for a crime. Second, partly as a result of the short time they are expected to devote to this task, they complain about the subjectivity and poor accuracy of their estimations. Accordingly, they hope that the accuracy of their predictions can be both improved and made more objective, or, at least, uniform. Based on an internal poll, only a minority of these judges cares about the logic behind the risk assessment being made accessible to them. The majority does not consider this essential as long as there is strong evidence that the tool is accurate.

2) Duty cantonal lawyers: the duty cantonal (defense) lawyers do not welcome this move. They complain about the opaqueness of the tool, which, they claim, makes it impossible for them to defend the rights of the people they assist. They plan to take legal action if the algorithm behind the score is not made accessible—and they are ready to appeal against corresponding decisions up to the highest court of justice."

2.9 What protocols are in place to deal with system errors and malfunction?

[Here you include a section that considers cybersecurity. E.g., you explain how you deal with errors by employees that compromise the integrity, availability, or confidentiality of the information collected. Optimally, this section is drafted by cybersecurity experts.]

2.10 What methodologies have been used to define and protect privacy?

[Here you explain methodologies actually used to protect the personal data with respect to privacy.]

2.11 What cybersecurity measures are in place?

[In this section, you explain cybersecurity safeguards built into the system. Optimally, this section is drafted by cybersecurity experts.]

2.12 What methodologies have been used to define and measure the bias and fairness of the system?

Example: "We have measured the false discovery rate, false omission rate, false positive rate and false negative rate in the aggregate as well as disaggregated by gender, assuming judges will deny parole to any defendant with a risk score higher than 0.3. However, the false omission and false discovery rates are equal for both genders. Risk scores have the same predictive value (they offer equally strong evidence that a parolee will reoffend) irrespective of the gender for which they are used.

The system is not designed to be adaptive, because an adaptive system would not have a homogeneous output for individuals with the same features. That is, people with the same characteristics may receive different recommendations, in a way that would be difficult to control by the judges."

2.13 How are individual predictions/recommendations/decisions of the system explained to system end-users and individuals affected by the use of the system?

Example: "The judges have been informed of the features that are considered by the prediction tool. However, the formula and the respective weight of these features are a trade secret and for this reason, they have not been revealed to them."

2.14 Is system deployment continuously monitored after the testing phase …

a) at all times?
b) within a given timeframe?
c) through which measures?

"No. No monitoring system is in place to track the performance of the system over time."

2.15 Are there options for people affected by a decision to learn about the output of the automated system and to challenge predictions/recommendations/decisions influenced by the system?

Example: "No, no such system has been designed and put into operation."

d) Transparency information on performance of the system

According to the tests run so far:

2.16 How does the system perform relative to the chosen relevant metrics?

[Please refer to all the goals and desiderata described when answering checklist 2—*questions 2.1 and 2.2*]

"The primary goal is to make predictions as accurate as possible: The system should minimize the number of people who are denied parole even though they would not reoffend as well as the number of those who are released on parole but who would reoffend. The system has an overall accuracy rate of x%. Its comparison with relevant national and international benchmarks is described in section 2.17 below."

2.17 How does the system compare to the one previously in place, if any, or to established benchmarks, if available?

Example: "The system previously in place was a parole decision by human judges, taking into consideration X, Y, Z. Unfortunately, there are no reliable studies about the overall accuracy of parole decisions by Swiss human judges (i.e., about the proportion of people released on parole who reoffend). There are some studies in the US showing that predictive tools used in the US are x% more accurate than human judges, but it is unclear whether and how this transfers to the Swiss context, since the performance of Swiss judges is not known. To measure the accuracy of Swiss COMPAS, we assume that a recommendation to deny parole is taken for any defendant with a Swiss COMPAS risk score higher than 0.3 regardless of gender (i.e., if this threshold is used, out of 1000 people released on parole, approximately 300 would reoffend). With this threshold, the Swiss COMPAS tool has an accuracy of x%, which is lower than the accuracy of the US COMPAS tool, which is y%. However, the false positive rate (the proportion of people predicted to reoffend in the test data who have not actually reoffended) is much lower in Swiss COMPAS than US COMPAS, while the false negative rate is slightly higher. Hence, compared to its US counterpart, Swiss COMPAS performs better in identifying people who would not reoffend and deserve to obtain parole, but it performs worse in identifying people who reoffend and should be kept in prison."

2.18 What are the remaining security and privacy risks and why are they reasonable?

[Here you explain why the cybersecurity risk resulting from 2.9 and 2.11 is considered reasonable given the stakes and probabilities of an attack.]

2.19 Please describe relevant unresolved biases or possible sources of unfairness in the system and explain why they cannot be resolved (e.g.,

by explaining trade-offs with other system goals, including conflicting fairness goals).

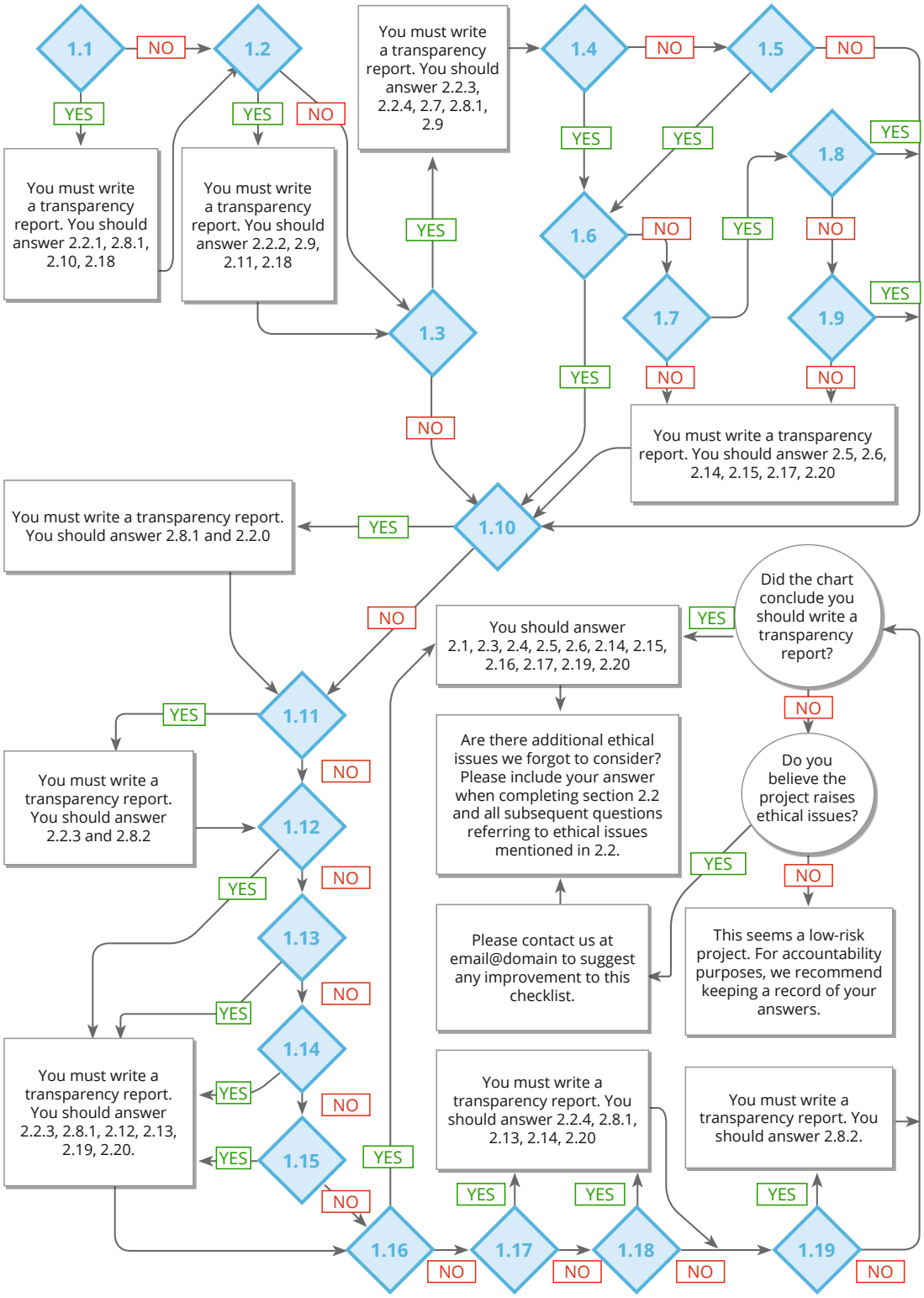
Example: "It is impossible to equalize all above-mentioned rates for all groups, because average rates of recidivism differ for men and women. For this threshold value, the false positive rate and the false negative rate differ with respect to different genders. Female prisoners have a higher false negative rate but a lower false positive rate, i.e., they are less likely to be wrongly detained and more likely to be wrongly released compared to men. Male prisoners have a higher false positive rate and a lower false negative rate, i.e., they are more likely to be wrongly kept in prison and less likely to be wrongly detained.

Given the fairness goals specified in 2.2.3 and the importance of equalizing the false discovery and false omission rate (as well as communicating calibrated scores to judges), which follows from this analysis of the goals of the system, we contend that this solution scores best from the point of view of fairness."

2.20 During monitoring, have predictions/recommendations/decisions by the system ever been challenged?

Example: "The system is not yet in operation in Switzerland. It is possible that similar systems may have been challenged in other countries in which they are deployed, but our department does not possess this information."

5 Flow Chart


1.1
NO
1.2
YES
YES
NO
You must write a transparency report. You should answer 2.2.1, 2.8.1, 2.10, 2.18
You must write a transparency report. You should answer 2.2.2, 2.9, 2.11, 2.18
You must write a transparency report. You should answer 2.2.3, 2.2.4, 2.7, 2.8.1, 2.9
1.4
NO
1.5
NO
YES
YES
1.8
YES
1.6
NO
YES
NO
1.7
1.9
YES
NO
NO
1.3
YES
NO
YES
NO
You must write a transparency report. You should answer 2.5, 2.6, 2.14, 2.15, 2.17, 2.20
You must write a transparency report. You should answer 2.8.1 and 2.2.0
YES
1.10
NO
Did the chart conclude you should write a transparency report?
YES
You should answer 2.1, 2.3, 2.4, 2.5, 2.6, 2.14, 2.15, 2.16, 2.17, 2.19, 2.20
NO
1.11
YES
NO
Are there additional ethical issues we forgot to consider? Please include your answer when completing section 2.2 and all subsequent questions referring to ethical issues mentioned in 2.2.
Do you believe the project raises ethical issues?
You must write a transparency report. You should answer 2.2.3 and 2.8.2
1.12
YES
NO
YES
NO
Please contact us at email@domain to suggest any improvement to this checklist.
This seems a low-risk project. For accountability purposes, we recommend keeping a record of your answers.
1.13
YES
NO
1.14
YES
NO
You must write a transparency report. You should answer 2.2.3, 2.8.1, 2.12, 2.13, 2.19, 2.20.
YES
1.15
YES
NO
YES
You must write a transparency report. You should answer 2.2.4, 2.8.1, 2.13, 2.14, 2.20
You must write a transparency report. You should answer 2.8.2.
YES
YES
YES
1.16
NO
1.17
NO
1.18
NO
1.19
NO

Automated Decision-Making Systems in the Public Sector: An Impact Assessment Tool for Public Authorities

Michele Loi, in collaboration with Anna Mätzener, Angela Müller, and Matthias Spielkamp

June 2021

Available online at https://algorithmwatch.org/en/adms-impact-assessment-public-sector-algorithmwatch/ and https://algorithmwatch.ch/en/adms-impact-assessment-public-sector-algorithmwatch/

Publishers:

AW AlgorithmWatch gGmbH
Linienstr. 13
10178 Berlin
Germany
https://algorithmwatch.org

and

AlgorithmWatch Switzerland
Spindelstr. 2
8041 Zürich
Switzerland
https://algorithmwatch.ch

Contact: info@algorithmwatch.org and info@algorithmwatch.ch

Copy editing:
Graham Holliday

Layout:
Beate Autering, Beate Stangl,
www.beworx.de

ALGORITHM WATCH

ALGORITHM WATCH / CH