Philosophy & Technology (2022) 35: 33
https://doi.org/10.1007/s13347-022-00528-0

RESEARCH ARTICLE

Promises and Pitfalls of Algorithm Use by State Authorities

Maryam Amir Haeri[1] · Kathrin Hartmann[2] · Jürgen Sirsch[3] · Georg Wenzelburger[2] · Katharina A. Zweig[2]

Received: 18 June 2021 / Accepted: 24 March 2022 / Published online: 9 April 2022

Abstract
Algorithmic systems are increasingly used by state agencies to inform decisions about humans. They produce scores on risks of recidivism in criminal justice, indicate the probability for a job seeker to find a job in the labor market, or calculate whether an applicant should get access to a certain university program. In this contribution, we take an interdisciplinary perspective, provide a bird's eye view of the different key decisions that are to be taken when state actors decide to use an algorithmic system, and illustrate these decisions with empirical examples from case studies. Building on these insights, we discuss the main pitfalls and promises of the use of algorithmic system by the state and focus on four levels: The most basic question whether an algorithmic system should be used at all, the regulation and governance of the system, issues of algorithm design, and, finally, questions related to the implementation of the system on the ground and the human–machine-interaction that comes with it. Based on our assessment of the advantages and challenges that arise at each of these levels, we propose a set of crucial questions to be asked when such intricate matters are addressed.

Keywords Algorithmic governance · Regulation of AI · Ethics · Accountability · Human–machine interaction

1 Introduction

Algorithmic systems have found their way into many parts of our everyday lives. They are selecting which articles will appear in our Facebook newsfeed; they are proposing Amazon products and are indicating to banks whether we are creditworthy enough to be eligible for a loan to buy a new house. Moreover, algorithmic decision-making (ADM) systems are not only part and parcel of our daily interactions with the private sector, but they have also increasingly been employed by state

✉ Georg Wenzelburger
georg.wenzelburger@sowi.uni-kl.de

Extended author information available on the last page of the article

authorities in recent years. The police has, for instance, been using algorithms to forecast where and when crimes will be committed in certain areas in order to more efficiently deploy police forces (Bennett Moses & Chan, 2018; Bowers et al., 2004); Social workers are exploiting information from algorithm-based tools to inform their work with children and families (Gillingham, 2019), and decision-makers in the criminal justice system are incorporating information about recidivism risks calculated by algorithmic tools when they make decisions about pretrial release of suspects (Hartmann & Wenzelburger, 2020).

These developments have raised questions about "algorithmic governance" to the top of the political and academic agenda (Danaher et al., 2017; Gritsenko & Wood 2020a, 2020b; Yeung, 2018a, 2018b) — and the debate on whether the freedom of the market should prevail and to what extent state policy action ought to provide a framework regulating such technological advances has been lively and included different disciplinary perspectives. Research on socio-technical systems (Beer, 2017; Holton & Boyd, 2020) has for instance discussed the very general question of how the advent of ADM systems transforms interactions between humans and technology. Still from a macro-perspective, legal-ethical scholars and political scientists have analyzed different forms of algorithmic governance (Gritsenko & Wood, 2020a, 2020b; Yeung, 2018a, 2018b) and their implications for social ordering (Katzenbach & Ulbricht, 2019) and democracy (König & Wenzelburger, 2020). Moreover, building on philosophical and legal arguments as well as design aspects, interdisciplinary research has extensively discussed how issues of fairness, transparency, and accountability can be dealt with when ADM systems are designed (Lepri et al., 2018; Mittelstadt et al., 2016; Tsamados et al., 2021).

However, while all these strands of the literature from computer science, ethics, law, public administration, and political science have contributed to a vibrant interdisciplinary field of research and enriched our understanding of both the technical side of algorithmic tools and questions of algorithmic governance, two aspects have been underrated in the current state of the art. First, while interdisciplinary work between individual disciplines (e.g., between ethics and computer science) does exist, studies that focus on the "big picture", namely, the key questions that need to be asked when states discuss whether to adopt an ADM system in a certain field of application, are only seldomly tackled in an overarching manner. Second, most of the work is conceptual in nature and discusses, for instance, ethical principles and possible operationalizations in data science, or governance principles and their consequences for public administration. However, empirical evidence from systems that are already implemented by the state and assist decision-making in the daily work of bureaucrats is only rarely included in these debates.

In this article, we therefore take a step back and look at these developments from a birds-eye perspective while, at the same time, tying together conceptual arguments from the literature with empirical traces from case studies. In this way, we aim at undertaking an inter-disciplinary discussion and integrating insights from computer science, politics, and public administration as well as ethics to identify key aspects of how the relationship between the state and its citizens is affected by the advent of algorithms. We focus on the state, because it is clear that while the applications themselves are similar whether used by the private sector or by the state, the use of

ADM systems by public authorities is nevertheless more critical. In this case, citizens are not users of these systems but are subjected to their results. This implies that they often do not have the liberty to opt out of being treated by an ADM system, and they do not have any direct contact with the output of these systems. Also, state action has to be exemplary in terms of how it obeys the rule of law and transparency criteria. Hence, if state authorities use ADM systems, their actions have to live up to the highest standards in terms of the protection of civil rights, individual freedom, and liberty.

Our key concern in this paper therefore is to analyze the key questions to consider when public authorities think about buying and employing an ADM system in a certain area of decision-making. To do this, after having briefly sketched out the main issues of the state of the art, we draw up a process model that includes the main steps to be taken in such decisions (Sect. 2). Based on this model, we discuss each of the key steps including ethical and political perspectives as well as insights from data science. In our concluding section, we discuss the implications of our analysis for political and administrative decision-making about the implementation and the use of algorithms.

2 State of the Art and Analytical Framework

2.1 Algorithms in Politics and Society: a Brief Review of the State of the Art

In recent years, we have seen a growing body of research discussing the consequences and challenges that arise with the increased use of algorithmic systems throughout society (for recent overviews from somewhat different perspectives, see, e.g., Yeung & Lodge, 2019; Barfield, 2020; Tsamados et al., 2022 or Mitchell et al., 2021). However, while we have gained important insights from this increasing number of studies, the state of the literature has to cope with two challenges: First, as the field of research is expanding in various directions at the same time — mainly because scholars from different theoretical and disciplinary perspectives such as ethics, computer science, law, or public administration have become interested in the changes brought about by the rise of ADM systems — the literature is becoming increasingly disparate. In fact, only rarely are the different perspectives integrated into an overarching framework including for instance technical, philosophical, political, and administrative perspectives. And second, most of the studies remain primarily conceptual in nature and are short of empirical evidence from the actual implementation experiences. This is partly due to the fact that real-life applications by public administrations have only started to emerge in the last decade or so, but there also seems a certain disconnect between conceptual work on the one hand and empirical studies looking at implementation on the other. In this brief review of the literature, we will first try to systematize three main bodies of the literature[1] in order to set the stage for discussing the contributions this paper seeks to make.

[1] We propose a rather general overview of the literature here citing some of the pertinent work. We do acknowledge that there are many field-specific articles that discuss for instance ethical issues in certain settings such as medical applications, criminal justice, allocation of job training for the unemployed, admission to university, autonomous driving, and so on (Grote & Berens, 2020; Hudson, 2017; Lepri et al., 2018), but also technical solutions to intricate design problems or challenges that arise for in public administration.

A first important strand in the literature deals with the question of how to evaluate and design fair, accountable, and transparent algorithms. There is an important ethical debate about these issues (see for instance the reviews by Blacklaws, 2018; Lepri et al., 2018; Franke, 2021), although questions of fairness and transparency can only be one aspect of a more fully fledged ethical analysis (Mittelstadt et al., 2016; Tsamados et al., 2022). Moreover, and importantly, the ethical arguments have been accompanied by and sometimes integrated in more "technical" literature on how to design ADM systems in order to address issues of fairness and discrimination, and significant work has been produced that links ethical considerations to data science applications (Berk et al., 2018; Heidari et al., 2019; Segal et al., 2021). An important discussion concerns for instance the applicability of certain principles of justice, such as fair equality of opportunity (Rawls, 1999), to derive criteria for the design of algorithms or entire decision-making systems (Franke, 2021; Joseph et al., 2017; Lee et al., 2021; Noriega-Campero et al., 2019; Shah et al., 2021). Still, some argue that due to limitations of these approaches and trade-offs with other relevant values such as efficiency, there is also a need to design procedures for employing and designing an ADM system in a democratic manner by setting up deliberative procedures which ensure that all relevant interests — especially those of vulnerable groups — are taken into account (Donia & Shaw, 2021; Robertson et al., 2021; Wong, 2020). However, while some of this work bridges the disciplinary boundaries of ethics, computer science, and political theory, legal aspects, public administration issues, and empirical evidence from frontline implementation experiences are only rarely included.

A second body of the literature relates to work by legal scholars, researchers of public administration, and administrative law who discuss how the advent of algorithms challenges the inner workings of public administration (Kim et al., 2021; Lodge & Mennicken, 2019; Yeung & Lodge, 2019; Young et al., 2019). They often use theoretical perspectives of their disciplines — such as governance (Danaher et al., 2017; Gritsenko & Wood, 2020a, 2020b; Katzenbach & Ulbricht, 2019) or (risk) regulation (Ulbricht & Yeung, 2022; Yeung & Lodge, 2019; Yeung, 2018a, 2018b) and also include considerations from ethics or human rights to analyze how far the advent of algorithms has affected standard working principles and the rules of the games established in law (coining new concepts such as algorithmic regulation or algorithmic governance). An important question in this connection is how agency and the legitimacy of decision-making by state authorities has been altered (Busuioc, 2021; Zouridis et al., 2020), which is why a discussion of the limitations and possible solutions to address the changes brought about by an increased reliance on evidence generated by algorithms has been a core topic in this literature (Bullock, 2019; König & Wenzelburger, 2021; Krafft et al., 2020).

Thirdly, on a more general level, scholars have, finally, sought to identify some core principles that should be followed when implementing algorithmic systems in democratic decision-making processes. These principles have been discussed in relation to ethical considerations, such as the requirement that decisions by state authorities that affect individuals substantially must be contestable in order to prevent individuals from being dominated by the state (Pettit, 1997). In the literature on ADMs, this requirement is often (at least implicitly) interpreted as implying a right

of those subject to decisions by ADMs to "obtain human intervention, to express a view and to contest the decision" (Blacklaws, 2018: 3). In addition, the employment of algorithms raises questions of moral and legal responsibility — are those who design algorithms responsible for outcomes related to the employment of these algorithms (Martin, 2018)? Political scientists have also contributed to this more general debate pointing out the challenges that arise for the input side (Habermas, 2021), the throughput and the output-side (König & Wenzelburger, 2020) of democracy when algorithms are increasingly used in the public sphere.

However, while the growing literature has indeed dealt with many important issues and also involves interdisciplinary work, proposals aiming at a synthetic view on the key questions to be addressed when state authorities decide about adopting an algorithmic system have mostly been put forward by think tanks, Parliaments, or expert bodies. For example, the British Academy and The Royal Society (2017: 51) propose quite general principles such as the promotion of "human flourishing," which also include the protection of "individual collective rights and interests," requirements of accountability, inclusive and democratic decision-making, and institutionalizing systematic learning "from success and failure." Similar principles have also been discussed in the British House of Lords report on Artificial intelligence, although technological and economic aspects are more strongly articulated (House of Lords, 2018). And most recently, the European Commission has put forward a proposal for European regulation of AI (European Commission, 2020) — a work that was prepared by the high-level expert group of the Council of Europe which recommended several general guiding principles for the use of algorithms from a human rights perspective (Council of Europe, 2019). Undeniably, these works are important milestones as they show the intricate choices policy-makers are confronted with when deciding about whether to use an algorithmic system in a certain area of application and how to regulate it. But, as much of the work discussed before, they often remain on a general level of policy recommendations and do not strongly engage with empirical evidence from real-life examples of existing applications of algorithmic systems in everyday decision-making situations or from the messy reality of data analysis.[2]

Against this background, this paper has two main aims. First, we want to draw together arguments from the different strands of the literature as briefly sketched out before — with insights from ethics about general principles, from data science on modelling and design choices, and from legal studies and public administration on, for instance, questions of regulatory choices. Second, we want to include empirical evidence from policy studies that have looked at the on-the-ground implementation of algorithms for decisions of public authorities. The paper seeks to tie these parts together by focusing on some key decisions that a state necessarily needs to address when deciding about the adoption of an ADM system. The next section sets up a process model of such key decisions that will guide the subsequent analysis.

[2] While the European Commission's White paper does tackle issues of data (e.g., EC, 2020, 18–19), the guidelines remain (necessarily) on a rather general level.

2.2 Key Decisions when Using ADM Systems by the State

When a public authority ponders the question to use ADM systems in a certain area of application, case study research shows that such decisions are often taken at the lower levels of public administration. The HART-model of predictive policing was for instance developed within Durham Constabulary in UK in a bottom-up manner (Oswald et al., 2018). Similar examples have been reported from the USA, e.g., in the realm of child protection (Chouldechova et al., 2018), and research on ADM systems in criminal justice has revealed a rather uncontrolled and unsystematic growth of ADM systems, which have often been implemented on the level of counties first, and led to a patchwork of different systems at work, before some steps toward a more standardized state-wide approach were taken (Harris et al., 2019; König & Krafft, 2020).

Whereas this pattern is — from a policy research perspective — not entirely surprising, as the bulk of new policies is usually elaborated in "policy subsystems" in which experts discuss such issues (Baumgartner & Jones, 1991) and does not necessarily involve high-level decision-making, it is nevertheless important to acknowledge that such a disordered bottom-up implementation of ADM systems in many parts of public administration may harm their prospects in the long run. If key decisions are taken without appropriate time and expertise or while the system is already up and running, critiques for certain weaknesses or malfunctioning can actually damage the entire system to an extent that it will not survive, due to political reasons (see, e.g., Wenzelburger & Hartmann, 2021). In what follows, we therefore sketch out a stylized model of steps that are ideally involved when state authorities plan to introduce an ADM system in a certain area of decision-making (see Fig. 1). It will enable us to provide a systematic discussion of the challenges that arise at each step and the intricate choices that arise for the involved actors at each step.

The very *first aspect* to be tackled is the question whether an ADM system should be used in a certain area of application at all. This comes down to the question whether we can draw certain "red lines" and delineate areas where ADM systems should simply not be used. Such red lines could concern normative and ethical questions, but also relate to legal foundations or technical issues — for instance, if valid data is simply not available for training an ADM system or if there is no valid ground truth to be gathered (e.g., if the outcome is biased by the intervention itself). Clearly, such decisions are political in nature and will be taken by elected officials, advised by bureaucrats and expert bodies including ethics committees that can evaluate the intricacies of such questions.

If, however, the use of an ADM system is not generally ruled out, the question of *how to regulate its use* in the respective decision-making context comes to the fore (*stage 2*). This is where much of the work on algorithmic regulation as well as many of the proposals put forward by expert bodies come into play, as it is up to the political sphere to set up a regulative framework which describes the legal basis and the restrictions for the use of ADM systems in a certain area of application. These guidelines may involve certain implementation rules (e.g., whether the human that interprets the ADM output can overrule the decision) but also relate to more technical issues, such as instances where ADM output should

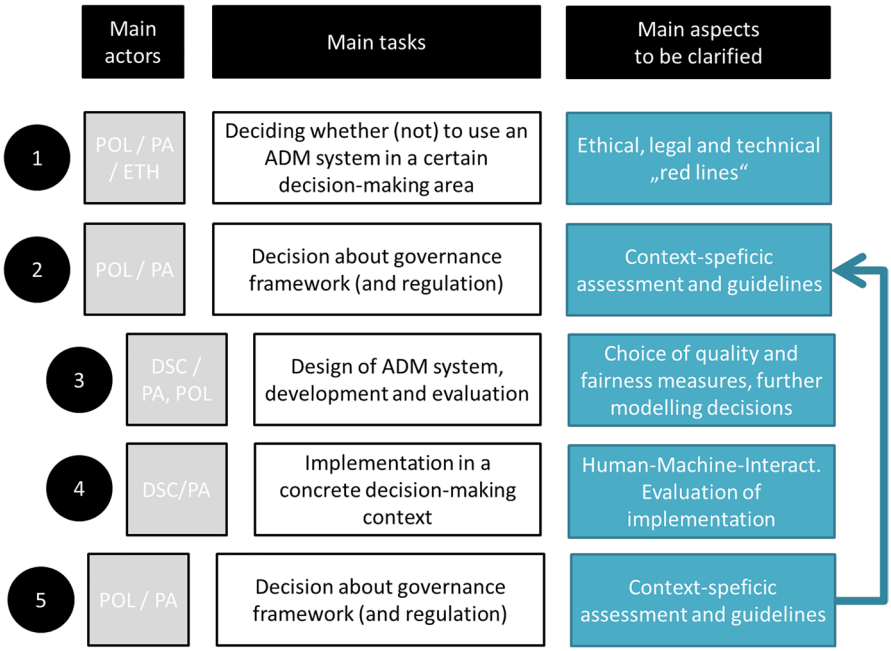

Fig. 1 Stages in a decision process leading to introduction of an ADM. ETH ethics; DSC data science; PA public administration; POL politics

not be used for certain categories of persons, because the system does not work well for such groups. As the arrow on the right in Fig. 1 shows, these decisions are interlinked with feedback from the two further steps in the process, namely, insights from the development stage on the design choices of the algorithm and experience from the concrete implementation of the system. Therefore, we have taken up step 2 again as a final fifth step in the process — and the two examples given above clearly show why this feedback loop is crucial: The regulation of implementation depends on both an evaluation of how human deciders interact with the ADM system and regulative decisions concerning technical issues on information about the performance of the system for certain categories of data.

The next key decisions concern the *design of the ADM system*, such as the choice of a quality measure, which indicates how well an ADM system predicts the "ground truth" given in a learning dataset. Moreover, fairness measures are also key in this step, because if ADM systems categorize people based on their characteristics (e.g., where they live, how old they are), this always involves the questions whether these groups are treated in a fair manner — a question on which much has been written (see the review section, above).

Finally, a crucial step — which is often overlooked — concerns the *concrete implementation of the ADM system* in what can be called the socio-technical environment. As research on the deployment of ADM systems shows, it is not enough to focus on the systems themselves, as even a well-designed system can be used by the

humans in a way that leads to problematic decision outcomes (Hartmann & Wenzelburger, 2020; van der Voort et al., 2019). The key task in this step is to conduct valid socio-technical evaluation studies on the implementation of ADM systems, which then can be used to improve the implementation on the ground, but can also lead to an adjustment of the regulatory framework (step 2) at a higher level if the evidence from the implementation studies points to substantial problems at this stage (see the feedback loop arrow to the right).

3 Analysis

Based on the heuristic sketched out above, the following sections will (1) discuss the challenges that arise for decision-makers at each of the stages and (2) give empirical illustrations from research on how ADM systems have been introduced in real-life settings. The empirical illustrations are mainly based on three case studies on the use of the COMPAS risk assessment system in the Criminal Justice system in Eau Claire County in the USA, the roll-out of the university admission system APB in France, and the plans to introduce the AMAS-system in the Austrian labor market agency.

3.1 Red Lines: Constraints by Data, Legal Frameworks and for Ethical Reasons

The very first decision to be taken concerns the question whether there are any arguments that would rule out the use of any ADM system completely, which would then mean to stick with human decision-making. There are several aspects that may lead state authorities to come to the radical decision, not to use an ADM.

The first point refers to *data*. In some fields of application, the data is simply not suited for the use of ADM systems — either because the training data the ADM learns from is severely biased or because there are systematic reasons, e.g., the absence of a valid ground truth. The applications in risk assessments in the criminal justice are a case in point, here. On the first argument, it is well known that there is a systematic and high underreporting of criminal offences (Arial & Bland, 2019; Killias et al., 2010: 20–28). The offences which pop up in the police statistics are severely biased toward blue-collar crime, violent crime in public, and property crime. Instead, criminal offences that take place within a household (e.g., a man beating up his wife) are underreported and so are white-collar crimes (e.g., UK Statistics Authority, 2014). However, as the data on which an algorithm is trained to predict risk is based on information about offenders that have effectively been caught (and reported to the police), the characteristics of underreported offenders are not sufficiently represented in the prediction. Consequently, the typical features which are statistically significant predictors of recidivism risk are those linked to the reported crimes, but not to the non-reported ones.

The second point is more problematic still, because it is based on the systematic problem of "asymmetric feedback." In the criminal justice example, this problem arises if ADM systems are used to predict recidivism in order to inform decisions

about suspects: In such cases, we simply do not know what would have been the outcome if a person would not have been put in jail but let go free — that is, whether she would have recidivated. Hence, if the ADM bases its risk assessment on evidence from those cases that actually have recidivated after incarceration, it uses a systematically flawed ground truth. In fact, in order to build an ADM to assess the features of recidivism based on a big data training set (and assuming that underreporting is low), we would need to start a big experiment and let go free all suspects to see who would actually commit a new crime.[3] However, for ethical reasons, we would not want to conduct such an experiment, evidently.

Besides such important data-related issues, *ethical principles* are crucial in this first step, too. In fact, from a purely ethical perspective, one could also formulate normative objections against the use of ADM systems. Drawing on Rawls' theory of justice[4] (Rawls, 1999), it can be argued that statistical discrimination, which adversely affects disadvantaged groups, ought to be prohibited in societies where some groups suffer from structural disadvantages, because these disadvantages would likely be reproduced by the ADM. This is particularly problematic, if they are based on characteristics of individuals that are not shaped by these individuals themselves (e.g., where you grow up) and if the characteristics do not have a plausible direct causal connection to the predicted outcome. Moreover, the area of application matters, too. If the influence of an ADM system has strong repercussions on an individual's life chances, discrimination looms larger than in other, less consequential areas.

In criminal justice, for instance, the incarceration of individuals directly affects their fundamental basic rights and liberties. With respect to basic liberties, equal treatment in a strict sense has a much higher priority than for instance the allocation of advantageous positions by a recruitment algorithm. This is because the equal guarantee of these basic liberties should not be traded off for the sake of efficiency or other, less important goods (Rawls, 1999). However, this also means that a decision procedure akin to affirmative action, which could be implemented on the design level (see below), seems to be ruled out when it comes to sentencing decisions for individuals: As individuals have a right to be only evaluated with respect to legally relevant aspects of their actions (whether they have done something wrong) and with respect to the degree of control they have had over their actions, questions unrelated to the matter at hand such as socioeconomic background, group membership often used in risk assessment tools cannot be used in the decision (Angwin et al., 2016).

Finally, the *legal framework* is important, too. As Martini et al. (2020) discuss for Germany, constitutional rules can considerably restrict the applicability of ADM systems in certain areas. For the question whether a centralized algorithmic system for university admission could be introduced in Germany, the authors point to

[3] And even this might not be enough because we would assume that relevant statistical associations remain stable over time. Thus, we would have to re-conduct the experiment in certain intervals because ground truth might change over time.

[4] There is a growing literature on how Rawlsian principles could be effectively implemented in ADM systems (Franke 2021; Heidari et al., 2019).

several legal restrictions which can be partly tied back to the unchangeable first 20 articles of the German Basic law (Martini et al., 2020: 13, 17). Moreover, the European Data Protection Regulation also imposes legal barriers to the use of ADM systems (Martini, 2019) — partly because it rules out completely automated decisions without human interference (Vedder & Naudts, 2017), and partly because it asks for explainability of the decision produced with the help of ADM systems (Brkan, 2019).

While these regulations do vary between nation-states, some of the core points are based on general civil liberties and therefore provide comparable legal constraints in liberal democracies. This is for instance true in the realm of criminal justice, where the need to do justice to an individual is closely linked to the idea to assess the individual as a whole and not as a bunch of characteristics (Harkens et al., 2020: 24). If such principles are important in a legal framework, this may also rule out the application of ADM systems in a very early stage.

In sum, it is clear that all three aspects — data, ethical considerations, and legal constraints — may form "red lines" for the implementation of ADM systems by public authorities. They are powerful boundaries for the introduction of ADM systems, and public authorities would need to assess these aspects. In practice, this necessitates to set up expert bodies which give advice to political decision-makers — not only on legal constraints (such as the GDPR (e.g., Brkan, 2019)), but also on data-related issues and ethical questions. Especially the last point also involves a normative assessment of basic societal principles, and it is important to allow a profound discussion by ethical committees — just as it is the case in other intricate normative questions. This may also include measures of citizen involvement to tie the deliberation back to society. From a democratic point of view, it is clear that the final decision about these normative questions needs to lie with democratically legitimized decision-makers.

3.2 Context-Specific Regulation: Assessing Potential Harm and Agency Loss

Once the general decision to use an ADM system has been taken, the most important question for public authorities is to develop a mode of governance that fits the particular field of application. As for the basic question on the use of the ADM, much depends on the context in which the system will be implemented. Based on the theoretical idea of the principal-agent model, it seems sensible to decide between two main dimensions that structure the assessment of how ADM systems should be regulated (for a more complete elaboration, see Krafft et al., 2020): The extent of agency loss on the one hand and the degree to which a decision interferes with an individual's life chances on the other.

The first dimension, the *extent of agency loss*, comes down to the question to what extent the individuals affected by the ADM decisions are actually able to make informed choices about whether they want to be evaluated by an ADM system. This depends on three important aspects: (1) whether the person can actually stay clear of the ADM intervention, (2) whether she is informed about the inner workings of the ADM to a degree that she can openly challenge the outcome, and (3) whether the

decision is entirely automated or whether the ADM is only used to inform human decision.

On the first aspect, agency loss is particularly low in cases where individuals have a choice whether or not to use an ADM or to switch between systems. If job seekers can choose, for instance, whether their chances for re-integration ought to be evaluated by an ADM or by a human decider, agency loss is much lower than if there is no possibility to opt out. Many of the systems used by the state are constructed in a way that there is no such opting out possibility — not least because the authorities introduce these systems exactly because they want to provide standardized and efficient solutions (Allhutter et al., 2020). Additionally, in some cases, the requirement of equal treatment rules out choosing between two different processes. The introduction of the French system of university admission APB is a prime example of such reasoning. In fact, the introduction of a centralized system and the admission of students by algorithm was seen by the actors as preventing patronage and unfair decisions, because all candidates were treated in a similar way.[5] However, there are also examples where public authorities do give the affected individuals chances to opt out: When ADM-based risk assessment was introduced in the criminal justice system in Eau Claire County in the USA via the COMPAS system, suspects had for some time the right to decide whether they wanted a COMPAS to be made. However, it is also clear that once you have the possibility to use an ADM system that is seen as an "evidence-based" tool by the main human decision-makers (e.g., the judges), not to rely on this tool is rather the exception. In fact, in-depth research has shown that not doing a COMPAS would have raised questions, which is why it pretty quickly became the standard (and the opting out solution was eventually dropped).[6]

On the second point, it is evident that agency loss is decreased when individuals can at least evaluate the ADM performance in relation to the output it produces (and object to it). This is very clear in the case of the French APB-system, where students were furious when they did not manage to get an admission to a university program whereas other students did — for unexplainable reasons. In one of the interviews conducted with a senior bureaucrat in charge of APB, this became clear when he explained that in one particular year, two twins in one family graduated in the same year with a similar grade from high school, applied for the same university program — and one got admitted whereas the other did not.[7] Such outcomes were produced by a random draw which was implemented in the APB algorithm in case of ties, but this was not known to the applicants. Very strong agency loss also emerges when advanced ADM tools are used, which learn in unforeseeable ways and where an explanation of the decision process gets very complicated.

Finally, the degree to which decision-making processes are automated affects the extent of agency loss. If ADM systems are used to inform human decision-makers, agency loss is sometimes less palpable for those individuals affected by the decision compared to completely automated decisions. In fact, there is much room for

[5] Expert Interview, Paris, November 27, 2019.
[6] Expert interview Eau Claire, June 24, 2019.
[7] Expert interview, Lyon, March 2, 2020.

regulating the machine-human-interaction in the decision-making process — for instance in terms of the question when and how a human decision-maker is informed about the outcome of an ADM or how a human can overrule the advice given by an ADM. In the Austrian AMAS system, which assists street-level bureaucrats in assessing the job market prospects of unemployed, there are very clear guidelines indicating how a human can overrule the output of the AMAS system. Although, in this case, the hurdles are still rather high (the bureaucrat has to ask her superior whether she can overrule the system), it is important to clearly state how such processes are structured.[8]

The second dimension — *degree of interference in life chances* — crucially depends on the area of application. In criminal justice, for instance, decisions may touch upon basic rights such as individual liberty, which is why one could argue that ADM systems should be ruled out completely for ethical reasons (see above). In contrast, ADM systems used to calculate the probability to find a job in the labor market or to regulate admissions to universities affect individuals' life chances less strongly, while harm and agency loss are very limited in ADM systems that suggest fashion articles to consumers.

For the empirical cases which serve as illustration in this article — the adoption of the COMPAS risk assessment system, the APB tool for university admissions, and the AMAS profiling tool for the unemployed — the risk matrix outlined above suggests different extents of regulation. The COMPAS case, if not ruled out for crossing red lines, would necessitate particularly strict regulation as potential harm is very high and agency loss too (opting out possibilities are very limited and, at least in practice, not to be recommended). The case of the APB university admission tool falls in the medium category with a certain agency loss (because of its centralized setup and the lack of transparency of basic decision criteria) and a medium strong interference with life chances. The question of interference with life chances is intricate, though, because it could be felt very strongly by some individuals (if admission to the study program of one's dream did not materialize), but is rather limited in aggregate (APB consumer surveys indicated that most candidates were rather satisfied with the procedure). Regulating the system would — at least — aim for clarifying the objectives of the system and involve, for instance, a discussion about fairness criteria to be met and regular and transparent checks whether these fairness measures are met. The new system which has replaced APB in France, ParcourSup, did indeed try to achieve some progress in this direction with clear rules for affirmative action — that is, moving up applicants from lower income groups in the admission lists to account for unequal chances in society.[9] Similarly, regular and transparent checks of the performance of the system are also part of such accountability — such as the reports by the High Court of Finance in France on APB and ParcourSup (Cour des Comptes, 2017, 2018).

Finally, the AMAS case also falls in a medium category, due to the potential harm inflicted by the decisions of the system and a certain degree of agency loss.

[8] Internal AMAS documentation, expert interview, online, August 3, 2020.
[9] Expert interview, Paris, February 2, 2020.

Empirically, AMAS has been subject to some critiques (Allhutter et al., 2020; Berner & Schüll, 2020), but has also been cited as exemplary in terms of transparency and application guidelines (Klingel et al., 2020). Transparency on the algorithm has been created by the publication of extensive documentation by the designers of the AMAS system (Gamper et al., 2020; Holl et al., 2018), including the methods, the used quality measures, and some information about the implementation guidelines. However, while the official documents are explicit about the fact that AMAS is likely to be adjusted in future years, a systematic mechanism of feedback does not seem to exist. Finally, the decision about the criteria used in the AMAS algorithm has also been criticized (Lopez, 2020). Clearly, such decisions are difficult to take as they involve trade-offs between values, which is why a structured and continuing involvement of stakeholders is important (Lepri et al., 2018; Veale et al., 2018).

Summarizing the decisions to be taken at this second stage, it is therefore crucial for state authorities to assess with experts what degree of agency loss the use of an ADM system entails and how strongly it interferes with citizen's life chances in a certain area of application. Depending on this assessment, a systematic discussion of appropriate rules of governance of the system can be started resulting in a clear and transparent regulative framework. The most important actors in this stage are policy experts who can evaluate the application context best and political actors as well as senior bureaucrats to design an appropriate regulative framework. Moreover, extensive implication of stakeholders — especially those representing affected but disadvantaged groups in society — may be important from an ethical point of view (Jörke, 2013; Veale et al., 2018).

3.3 Development and Design Decisions

If an ADM system is to be implemented, the development and the design of the system is the most important element from a technical side. This step involves a series of very important questions that need to be answered:

- On data: What training data can be used to build the ADM? How good is data quality? How representative is the training data for the population about which a prediction is to be made?
- On model building: How should the model be built, i.e., are there causal relationships well identified in the scientific literature that can be used to build the model or will the data scientist explore the patterns in the data and come up with some model to predict the outcome of interest? Are there any variables that should not be used for building the model, because they may involve discrimination? What statistical method is appropriate for building the model?
- On model evaluation: What measures should be used to assess the quality of the model? What fairness measures are appropriate for the field of application? What criteria are to be met for a model to be "ready" for use?

Questions like these are often treated implicitly during the process of model building by data scientists and not sufficiently discussed on a more general level.

According to our interviews about the APB algorithm used for candidate selection in France, the data scientists were left alone with decisions about which additional criteria to use in order to avoid equally ranked candidates, because politicians did not want to tackle this intricate question of selection criteria.[10] This is, evidently, not an appropriate way to go forward. Instead, model building needs to be discussed before it is started, because every choice will affect the outcome and does involve normative and political trade-offs.

3.3.1 Nature of Data and Data Selection

Before model-building can start, the nature of the training data used to build the algorithm has to be scrutinized. Important questions in this context are, for instance, whether the training data is appropriate for the context in which the ADM will be applied, or whether the data quality of the training data is sufficient. On context, the COMPAS case is, again, very illustrative. As we have learned researching the implementation in a rural county in Wisconsin, they initially used the ADM trained with data from a densely populated area in California, before the algorithm was adjusted several months later.[11] Clearly, such context-blind application of a system is problematic.

Similarly, data quality can be a concern, if data points are missing, for instance, or if the reliability or the validity of the data is unclear due to absence of information about data generation processes. Using the example of criminal justice, again, under-reporting on crimes is a widely acknowledged phenomenon, which clearly leads to biased data from police records (Schwartz & Vega, 2017). Therefore, when building a model, data scientists have to be made familiar with the specificities of the data used in the field where the ADM system is to be applied. In fact, much insights from the debate in social sciences about the "data generation process" include important lessons about how to deal with empirical data (Uher, 2019) — and this may be especially relevant in cases where survey questionnaires are employed to collect data which will then feed into the ADM system.

3.3.2 Model Building

In the *model-building exercise*, data scientists fit models to predict the outcome of interest based on a training dataset. However, this process is also subject to important decisions that are not free of normative and political judgements. First of all, there can be ethical issues concerning the variables to be used for model building — e.g., for "directly discriminatory" variables related to race or ethnicity (Altman, 2011). This holds mainly for two reasons: First, it is unfair to be treated worse on the basis of morally arbitrary attributes, such as gender, ethnicity, and family background, for which one cannot possibly be held responsible (Cohen, 2009; Dworkin, 1981). Additionally, being treated on the basis of such characteristics implies that one is not being considered as an equal to other members of society that do not

[10] Expert interview, Lyon, March 2, 2020.
[11] Expert interview, Eau Claire, June 24, 2019.

possess these arbitrary characteristic(s) in question (Anderson, 1999; Rawls, 1999). This kind of treatment therefore involves a communicative and symbolic bad that undermines an individual's self-respect and communicates to others that individuals of this "type" are less respectable. It damages an important public good in liberal-democratic societies, namely, the assurance that everyone is considered as an equal member of society (Waldron, 2012).

Similarly, the use of proxies for membership in salient groups — or variables that correlate with membership in salient groups — must be avoided due to the danger of reproducing structural disadvantages for the most vulnerable groups in society. This can be a problem especially in societies that contain structurally disadvantaged groups, where group-members face disadvantages across various societal spheres (Esser, 1999; Jugov & Ypi, 2019: 7). There is a debate about such intricate decisions in the literature on fair algorithms (Matthew et al., 2016), but it is clear that such judgements should not be left to data scientists but must be addressed in interaction with ethical experts and on the political level taking into account the interests and views of vulnerable groups.

Interestingly, these considerations may overrule the prohibition of using characteristics of salient groups introduced above, when, for example, an ADM is built to counterbalance existing societal inequalities and introduces "affirmative action" which is intended to offset the disadvantages and discrimination that members of some groups experience (Dworkin, 1977; Fullinwider, 2018; Nagel, 1973; Thomson, 1973). Hence, including or excluding certain variables from model building necessarily needs to be discussed beforehand on a political level and informed by ethical reasoning.

Another key aspect to consider is the question of causality. In several fields of application of ADM systems, there is century-old empirical research on the outcome of interest. Hence, there also are solid empirical findings about possible explanations for certain outcomes. Take the question of unemployment as an example. Sociological and economic research on labor markets has produced millions of pages of research about job careers and possible factors that influence why some individuals more easily find a new job on the labor market than others (Caliendo et al., 2017; Granovetter, 1995; Heinz, 1999). The main question therefore is how data scientists include such evidence into their model building, or whether they simply search for patterns in available data in an exploratory way to find correlations that will then be used to build a model. To give an example, in the realm of profiling of unemployed, there has been a general discussion related to the question of how ADM tools aimed at profiling the unemployed actually relate to such scientific research (Caswell et al., 2010; Desiere et al., 2019).

Finally, the choice of the method to build the model is also not neutral. Evidently, an important question is the nature of the predicted outcome (e.g., whether it is binary, ordinal, or metric). But even if the level of measurement of a certain outcome is known, several methods can still be used, and each method has strength and weaknesses. For the example of pre-trial risk assessment in criminal justice, König, and Krafft (2020) show for instance that different techniques are used to predict the outcome of interest (mostly recidivism or failure to appear (or both)): Lasso regression, binary logistic regression, or some other forms of bivariate or multivariate

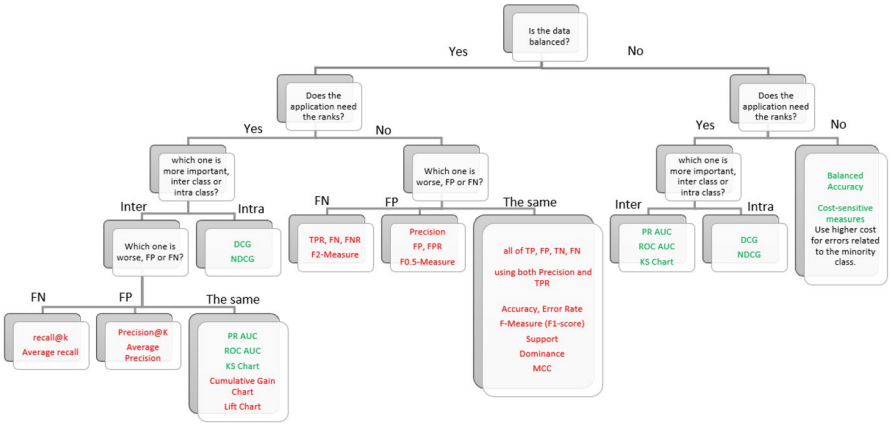

Fig. 2 Process model for selecting quality measure. Red text elements denote measures that can only be used for balanced data, whereas measures in green color can be used for both balanced and imbalanced data

analysis. While we do not know how the different methods have been chosen, it has been shown that method choice does have consequences in terms of modeling and prediction outcomes (Silberzahn et al., 2018).

3.3.3 Model Evaluation

Finally, the third important aspect to be covered when ADM systems are developed concerns model evaluation. This boils down to two main questions: The assessment of the quality of a system, i.e., how well does a system predict the outcome (compared to actual outcomes) and the assessment of fairness, which is particularly important when it comes to the classification of individuals based on group characteristics.

As different *quality measures* exist to assess the goodness of fit of a model, such as precision, recall, or the ROC AUC, it is important to choose the appropriate measure for the question at hand. In practice, to decide which quality measure is appropriate for evaluating an ADM with an underlying classification model, a stepwise process answering several questions is therefore sensible (see Fig. 2). The first criterion refers to the question whether the data is balanced. This is very important, because in the case of an un-balanced dataset, several popular evaluation measures, such as accuracy, do not reflect the quality of the system. In many applications, data is imbalanced such as in the area of criminal justice, because most citizens simply are not criminals. Thus, some quality measures may yield high scores even without classifying the minor class correctly.

The second question has to do with the output of the ADM, which can be a rank-order or binary. While ranks always can be transformed into binary values, this comes at a loss of information. However, for the selection of the appropriate quality measure, it is important whether the ADM outcomes are ranks or binary values. For example, consider a job hiring system that ranks the applicants and accepts

applicants with the highest ranks. In such a system, we need measures that consider the correctness of the rank orders and must choose the quality measure depending on whether inter-class or intra-class ranks are more important. In other words: Is it more important that the ranks of the instances belonging to the same class should be assigned in a well-ordered manner, or it is more crucial to rank the instances belonging to two different classes in a correct way? In the case of profiling job seekers, it may for instance be more important that the ranks of the individuals that are highly probable to find a new job on themselves are generally better than those of the applicants which need assistance; in contrast, in other fields of application, it may be crucial to focus on the rank-orders within one of the two classes. Whereas ROC-AUC [12] may be a good option if inter-class ranking is more important, DCG[13] is a better choice if one wants to focus on intra-class ranking.

In cases where the ADM uses binary class labels, the most important decision is whether false positive (FP) or false negative (FN) errors are more problematic. This question should be discussed beforehand in multi-disciplinary teams involving data scientists, ethics, and policy experts, not least because the decision whether to minimize FP or FN errors reflects on the quality measures to be chosen. For example, there is a trade-off between precision and recall: If the false positive error, which is reflected by precision, is more harmful in the domain of application, it is essential to evaluate the ADM system by such a measure. Thus, it is very important to identify which type of error is more important to make sure that the system works correctly. Besides, it is always recommended to evaluate the system with several measures especially those with trade-offs.

Besides the evaluation of the quality of ADM systems, it is essential to also consider the *fairness* of the systems. It is possible to have an ADM with high-quality decisions even though it is not fair. Moreover, similar to quality measures, there are many fairness measures defined for evaluating of ADM systems (for a recent overview, see Haeri & Zweig, 2020). However, the crucial challenge is that most of the fairness measures are mathematically in conflict with each other, because they operate with different fairness definitions. This, in turn, means that it is impossible to satisfy all mathematical fairness criteria at the same time and have a high fairness value based on all measures. Therefore, as for quality measures, we need to choose the appropriate measures for evaluating fairness in a certain context. The process model illustrated in Fig. 3 can be seen as a conceptual map for such decisions. It involves several questions, the answers to which need to be elaborated in multi-disciplinary teams and should be made not by computer scientists alone.

At the highest level, the model divides the available fairness measures into two main groups, those based on group fairness or distributive fairness (1) and those

[12] ROC-AUC is the abbreviation for "Receiver-operating characteristic — area under the curve". The ROC curve summarizes the performance of a binary classifier by comparing the true positive rate (recall) vs. the false positive rate at all possible values of the cutoff, and the area under the (ROC) curve indicates the overall quality of a binary classifier.

[13] The discounted cumulative gain (DCG) considers the in-class ranking. It does not only account for the fact whether a positive instance obtained a higher rank than a negative instance, but also differentiates between more or less relevant positive instances.

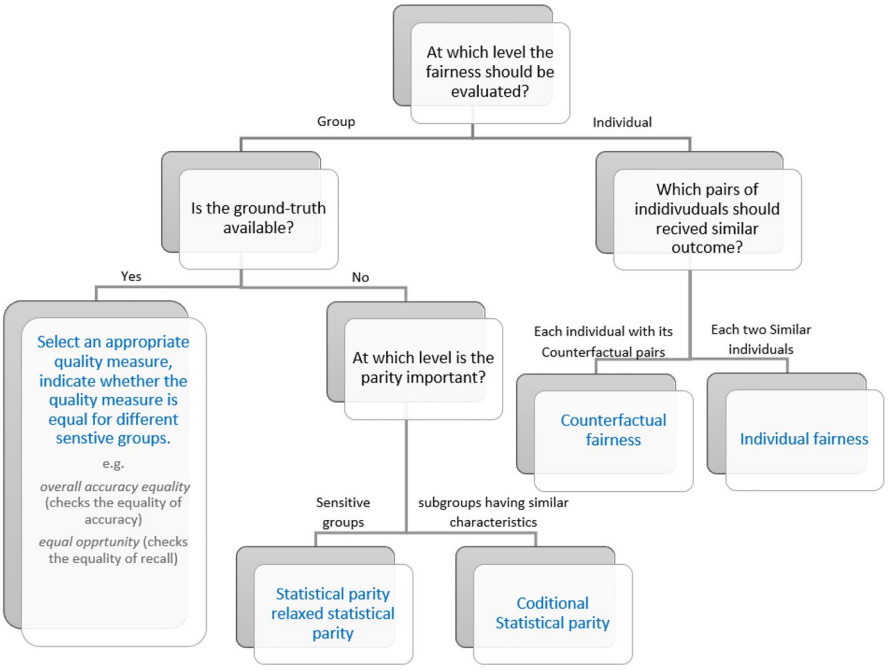

Fig. 3 Process model for selecting fairness measures

directed at individual fairness (2). Group fairness criteria ensure some sort of distributive parity for members of different relevant social groups, whereas individual fairness ensures that any pair of individuals who are similar should be treated similarly by the system. In general, group fairness measures are easier to implement and test. As most of the group fairness measures are based on quality measures, the procedure of selection proper fairness measures is closely associated with the process of choosing quality measures. In contrast, individual fairness criteria are more difficult to satisfy, and evaluation of individual fairness measures requires more computational costs. Thus, there is a trade-off between group fairness and individual fairness measures, which is why the first decision is to indicate at which level the fairness should be evaluated.

When it is decided to evaluate the fairness of an ADM system at the group level, the choice of the fairness measure depends on whether ground truth is available. Most of the fairness measures are based on the idea to compare quality measures across different sensitive groups — a procedure for which ground truth is required. Ideally, the quality measures should be equal for different sensitive groups. Hence, selecting the fairness measure can follow the same process as choosing the appropriate quality measure. We believe that in this situation, we should be consistent with the quality measure chosen for evaluating the quality of the ADM system. Then, we can evaluate the fairness of the system by checking the equality of the quality measure for different sensitive groups.

However, while ground truth is usually available for training the data in the development phase, this may change after the deployment of the system. Still, fairness measures can also be calculated after deployment without available ground truth for the real cases, simply by comparing outcomes for certain groups of individuals. The available measures to pick from can be categorized into two groups: statistical parity and conditional statistical parity. Which measure to choose depends crucially on the question whether it is more important that different sensitive groups have the same distribution of outcome (statistical parity), or whether it is more important that subgroups with similar profiles have the same distribution of outcome (conditional statistical parity). If statistical parity is key, then it is possible to use statistical demographic parity measures (depending on the sensitive categories, e.g., gender, age, ethnicity). However, if conditional statistical parity is to be reached, a proper measure for evaluating fairness depends on the conditions, by which subgroups of similar people are generated. For example, consider an ADM system to profile job seekers. If it is desired to have fairness in a way that the rate of accepted applicants for men and women is the same, then statistical parity measures can evaluate this type of fairness criteria. However, if it is important that the outcome of the system for women with specific skills is the same as those men with similar skills, then we need to use conditional statistical parity. Indicating conditions is a crucial procedure, and the conditions should be based on variables which are independent of the sensitive features. Again, this decision cannot be taken by data scientists alone, but in collaboration with ethics experts and policy specialists from the field of application.

When group fairness measures are not to be used, individual fairness measures can assess fairness at the individual level. The key question in this respect is which pair of individuals should receive a similar outcome. There are two major categories at this level — individual fairness and counterfactual fairness measures. Individual fairness measures indicate whether the output of the system is the same for similar individuals. Thus, they consider similar pairs of individuals and check whether the outcome of the system is the same for them. For this type of measure, a similarity measure is required to find similar individuals. This similarity measure is very important, and it should be chosen in a way that considers the merits of each individual, which are not dependent on sensitive features. The other category is counterfactual fairness. A system satisfies the counterfactual fairness criterion, if the system output for any member of a sensitive group (i.e., ethnicity, gender, sexual orientation) is the same as when he or she is from a different group. In counterfactual fairness, a causal graph between attributes is considered and then a counterfactual pair is created, consisting of one individual with certain attributes and another individual, who has the opposite value on the sensitive attribute but matches the first individual in all other features which are not proxy of the sensitive attribute. Based on such a pairing procedure, these measures evaluate the equality of outcome for the counterfactual pairs.

Thus, in the cases where fairness should be evaluated at the individual level, it is important to indicate which pair of individuals should receive similar outcomes — similar pairs, or counterfactual pairs? Generally, counterfactual fairness needs prior expert knowledge for assuming correct causal graphs between the attributes. If causal graphs are correct, new biases may arise. In the case of similar pairs,

individual fairness measures are dependent on defining a valid similarity measure. At any rate, such decisions cannot be left to data scientists alone, but need to be discussed in collaborative teams including ethics and policy experts.

3.4 Deployment and Implementation: Humans and Machines in Real Life

The last stage of our heuristic that describes the key steps when state authorities decide about introducing an ADM system concerns the implementation in the front-line work of state agents. These can be consultants of an employment agency using a profiling system, policemen using predictive policing tools or agents in the criminal justice system that use outputs of a risk assessment system to inform their decisions. As decisions of state authorities are — in most cases — not completely automated, but inform decisions by a human agent (the famous "human in the loop"), this last step involves the question how to organize the interaction between humans — e.g. street-level bureaucrats — and the ADM system in place.

Yet, before delving into the intricacies of human–machine interaction, the most basic question to be assessed on this level is whether ADM systems actually deliver decisions of higher quality than humans in a certain area of application or whether efficiency gains[14] can be observed. While this aspect does not touch on red-line decisions of stage 1, it could affect decisions on all other stages, especially if the ADM systems would not perform better than human deciders. In essence, one can treat this question as an empirical one, which needs to be assessed by comparing performance (in terms of quality or fairness), e.g., by means of experimental studies. One could, for instance, compare aggregate outcomes of performance of university students (during their studies or on the job market) that have previously been assigned to study programs by either an ADM or by human decision. If the ADM yields a better outcome, this would point to the value of its introduction. While such an insight enriches the discussion about regulation (stage 2) and introduction (stage 1), it is also clear that several other aspects will not be affected. Legal and ethical considerations are not influenced by performance: Even if an ADM system would yield a better quality of aggregate decisions, it could still not be ethical (from a fairness perspective) or legal to use it.

Most importantly, the introduction of an ADM is not a black-and-white decision. Instead, the human decision-making can also be assisted by the use of ADMs (Lepri et al., 2018), which brings us to the question of human–machine interaction. In fact, one can argue that the use of statistical evidence generated by an ADM system transforms the decision-making context in which bureaucrats take decisions (Hartmann & Wenzelburger, 2020). As Zavarsnik points out, "the process of arriving at a decision changes. The perception of accountability for the final decision changes too. The decision-makers will be inclined to tweak their own estimates of risk to match

[14] Moreover, it has to be evaluated whether the probable cost advantage of an ADM compared to human decision-making, especially if it involves many cases to be dealt with, looms large even if quality of human decision-making has been shown to be slightly better. These questions have to be dealt with on level 2, when decisions about regulation or governance will be discussed.

the model's" (Završnik, 2019: 13). Therefore, only if we know how ADM systems work on the ground can we assess chances and risks of ADM use (Veale et al., 2018). Two aspects are central in this respect.

First, as public agents are — as every human — keen on reducing uncertainty and ambiguity when they take decisions (Gajduschek, 2003: 715–717), the information provided by ADM systems is particularly welcome. With a score generated by an ADM output, human decision-makers, who had mostly relied on heuristics (Tversky & Kahneman, 1974; Vis, 2018) or on pragmatic abduction (Ansell & Boin, 2019) in their standard bureaucratic practices and followed established guidelines and rules (Bovens et al., 2014), can now base their decision on additional information. In fact, our research on the implementation of COMPAS in the criminal justice administration[15] reveals that such quantified evidence is often interpreted as "scientific objectivity" and "provides an answer to a moral demand for impartiality and fairness" (Porter, 1995: 8). This also means that the influence may be much bigger than mere "additional information" but serve as an important baseline which is difficult not to take into account or to overrule for a human agent.

However, we also find that the way in which the availability of algorithmic outputs changes the decision-making context varies much across contexts. For decisions in criminal justice, interviews with practitioners have indeed shown that the risk assessment tools were seen as very important sources of information and did change the decision-making process. This may be explained by the field of application, as decisions about whether a person will be held in custody to await trial or released carries to produce considerable harm — in case of actual recidivism (for the victim) and in case of jail (for the suspect). Hence, decision-makers may be very strongly inclined to use the evidence provided by the ADM, perhaps also to avoid blame for malign decisions. In other cases, such as the profiling of unemployed, our case studies on Austria's AMAS system also point to a certain reluctance to use the score produced by the algorithm to inform decisions[16] — and similar results have also been reported in other fields (Burton et al., 2020).

Second, what has transpired from all our case studies about frontline implementation of ADM systems, though, is the lack of adequate training measures for the agents of the state. While actors in the area of risk assessment in criminal justice received an initial training during the implementation process of the ADM, this procedure did not last for long. Instead, new actors that come into the criminal justice system received the information needed through studying documents and informal meetings with more experienced colleagues.[17] This practice however cannot be seen as a systematic and ongoing way to ensure that every ADM user holds a certain degree of digital literacy that is key in order to interpret the outcomes of ADM systems properly — especially when it comes to cases where the own assessment diverges from the output provided by the ADM. In contrast to that, a systematic and ongoing training approach was part of the implementation plan of the ADM used in the area of profiling of job seekers. However, persons in authority at the lower

[15] Expert interview, Eau Claire, June 24, 2019 and June 25, 2019.

[16] Expert interview, online, November 11, 2021.

[17] Expert Interview, June 24, 2019.

management level have been raising questions concerning the (1) capacities to fulfil the training schedule in time and (2) to fulfil them on a regular basis.[18] Therefore, in order to ensure that every actor is well informed, the establishment of an appropriate process to ensure digital literacy seems to be still needed.

From these insights, several lessons can be drawn. First of all, it is of utmost importance to include an assessment of the implementation of the ADM system on the ground to evaluate whether it is worthwhile introducing an algorithmic tool in a certain area of application. Without such an evaluation that takes into account the socio-technical system and human–machine interaction, no meaningful assessment of the performance of the ADM system can be made. Therefore, concrete decisions about algorithmic governance in a certain field of application can only be taken if this last step of ADM implementation has been evaluated properly. These assessments need to be done by external institutions and may involve qualitative techniques (observation, interviews) but also experimental studies on how humans interact in a certain setting with a machine. The results of such evaluations provide important information about how the interaction may be structured in certain decision contexts, e.g., at which point in time the information from the ADM will be given to a human actor or what steps a human actor needs to take to overrule the system. But it can also lead to adjustment on the design level, for instance on how much information will be given to a human and in what way (only very broad risk classes for possible recidivism vs. detailed information about the prediction).

Secondly, it seems crucial to empower frontline workers and their superiors in the public administration to exchange on ADM-related issues with data scientists. This involves training on both sides and would help to develop a common language of these different actors, which would improve the system in two ways. First, public agents would be empowered to assess outputs of an ADM on their own and to also interpret the outputs as an additional source of information (and not as a quasi-binding advice). And second, data scientists would be able to come back to bureaucrats when there are questions to deal with on a technical level.

Thirdly, the evaluation of the implementation stage is also highly important for the final decision about how an ADM system should be regulated in a certain area of application. If, for instance, our case study evidence according to which the characteristics of decisions in the criminal justice system are potentially so risky that the availability of statistical "evidence" produced by an ADM system looms large, this may entail that much stricter regulative standards have to be introduced. Hence, the feedback from both, the design as well as the implementation stage, are therefore crucial elements in the second stage where the mode of governance is decided upon.

4 Discussion and Policy Recommendations

Drawing on a simple heuristic of important stages in a decision-process leading to introduction of an ADM by state actors, we have discussed what key questions

[18] Expert Interview, November 2, 2020.

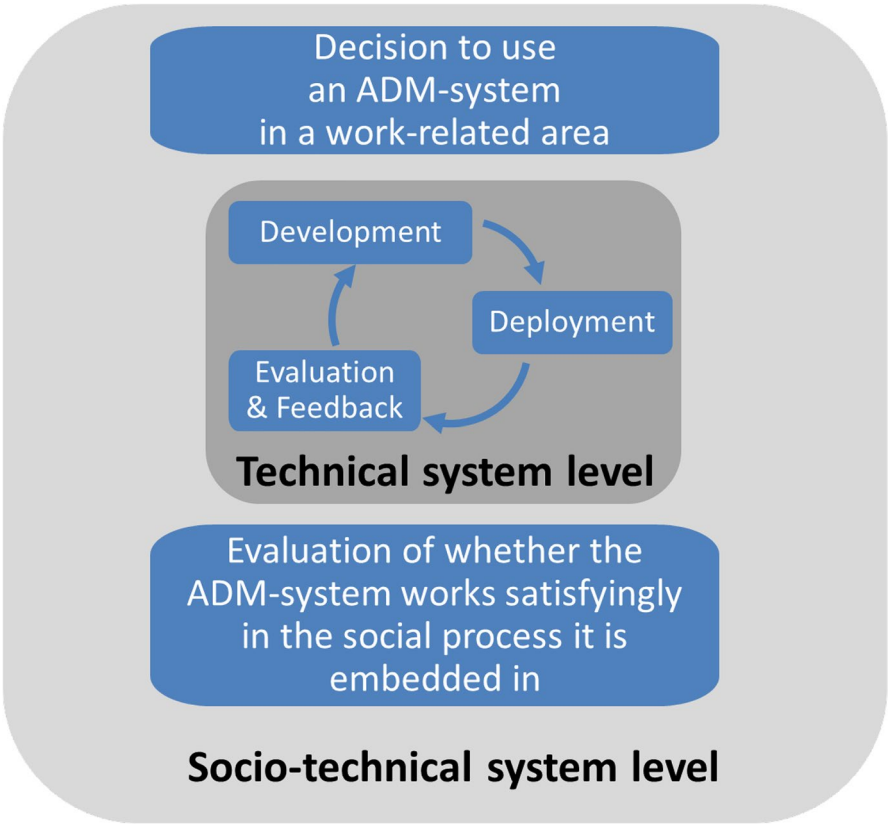

Fig. 4 A multi-layered account of ADM use by the state

may arise at each of these stages. In this concluding section, we will discuss what follows from our analysis in terms of general insights on the use of ADM systems by the state and in terms of possible questions that state actors ought to pose themselves when they ponder the question whether an ADM system should be used in a certain area of application.

On a more general level, our investigation emphasizes the need to conceptualize the use of ADM systems by state authorities as a multi-layered system that includes circular processes. As illustrated in Fig. 4, the use of ADM systems by the state does not only involve political and administrative decisions, but also a technical system level, which is usually outsourced to data scientists who design the system. Most importantly, these systems are interrelated and include two circular processes. On the technical system level, the development of an ADM system is affected by an evaluation of its performance, which then will lead to adjustments of the system itself. And on the socio-technical level, a similar feedback loop relates the overall evaluation of the implemented system in a field of application to the regulatory framework and the decision to use an ADM system.

Taking this argument seriously means, in practice, that there is the need to establish an obligatory and sufficiently extensive phase of experimentation and evaluation — both on the technical and the socio-technical level — before a final decision about the implementation of an ADM system can be taken by state authorities. Only then, the circular processes of evaluation and adjustment can be set in motion and inform the final decision, whether an ADM system should be used and how it should be regulated by the state depending on the context. Moreover, such a mandatory experimentation phase would also allow for stakeholders and citizens to be heard and involved in the process — especially when ethical considerations suggest that potential harm for disadvantaged groups may be high. Hence, feedback would not only be generated from the technical level and the level of the implementation by bureaucrats in the concrete field of application, but it would also include the views and opinions of important groups in society. How this process may look like can also differ depending on the concrete decision context (Lepri et al., 2018), but the general idea would follow the principle of giving a say to those groups in society that are likely to be most negatively affected by the introduction of the ADM.

Hence, our first and probably most far-reaching proposal for the decision of state authorities on the introduction of an ADM system is that an extensive phase of experimentation with ADM systems has to be mandatory for all implementations of an ADM by the state (see Table 1). This process should be overseen by an independent body and allow for an evaluation of the experimentation phase — with regard to technical issues and the implementation of the system in the socio-technical system.

Drawing on our discussion of the multiple stages presented above, several additional questions that need to be asked before the implementation of an ADM system by state authorities can be formulated (see Table 1). There are three key lessons that transpire from the summarized insights presented in Table 1. First of all, several key decisions can only be taken by democratically legitimized actors, that is, politicians that can be held accountable by means of the democratic process itself. This is especially true for the definition of red lines[19] as well as for the adjustment of the governance framework and the connected intensity of regulation. Clearly, advice from data scientists, ethics, and legal experts, as well as specialists from the policy area where the ADM is to be applied, is crucial in these decisions. However, the decision itself needs, again, related to the principles of the liberal democratic state.

Second, Table 1 also illustrates the necessity to set up multi-disciplinary teams to tackle some intricate problems, be it on the design level (e.g., on fairness measures) or on the implementation level (e.g., human–machine interaction). We are aware of the fact that this is a challenge and requires open-mindedness and collaboration over the disciplinary fields. Nevertheless, the continuing exchange on this level of expert knowledge is crucial to prepare decisions on the political level — especially in stage 2, where the insights from the design and technical development as well as the implementation phase come together.

[19] In many cases, some of the relevant red lines are already defined by the constitution.

Table 1 Guidelines for decision about and introduction of ADM systems by state authorities

Stage of decision making	Guidelines
(0) All stages	Before deploying an ADM system, an extensive phase of experimentation and evaluation ought to be introduced. The evaluation phase needs to be assessed by an independent body of experts and includes not only the technical aspects of ADM design but also the implementation stage
(1) Deciding whether to implement an ADM	Definition of red lines due to data quality, ethical principles, or legal issues. Preparation and advice by experts (ethics, legal, policy area), involvement of citizens, decision by democratically legitimized actors (e.g., committee)
(2) Deciding about governance framework	Assessing the intensity of regulation depending on degree of agency loss and interference of ADM in individual life chances in a particular field of application
	- Involvement of policy experts and stakeholders (especially of disadvantaged groups affected by ADM) in the process
	- Crucial: Introduction of dynamic feedback process during the experimentation phase and even after deployment
	- Decisions on this stage by democratically legitimized body (e.g., on questions of stage 3 and 4)
(3) Design and development	Discussion of data issues, modeling issues, and model evaluation
	- Collaboration of ethic experts and data scientists when it comes to selection of variables in model building (and questions of discrimination)
	- Setting up multi-disciplinary teams to evaluate fairness and quality measures
	- Decision about fairness and quality measures as key components of ADM design by democratically legitimized body (see stage 2)
(4) Implementation	Assessing human–machine interaction in real life context
	- Assess empirically whether the use of ADM systems produces better decisions in respect of the outcome of interest (e.g., experimental studies)
	- Study how decision-making processes change with the introduction of ADM systems (e.g., field experiments)
	- Empower frontline workers through continuing training measures delivered by independent bodies
	- Assure feedback of implementation stage in decisions about governance framework in experimental phase and once ADM is deployed for continuing adjustment (see stage 2)

Third, the summary of our insights again emphasizes the dynamic character of the decision on the introduction and the regulation of ADM systems. As ADM systems are based on training data from the past and as society as well as the context

of decision-making change continuously, this dynamic relationship has to be implemented from the outset in the governance framework. Take the current pandemic as an example: If labor markets are substantially changed due to a lockdown, how could then a profiling algorithm, which has, necessarily, never seen such a situation in the data on which it was trained, come to conclusive profiling results? In order to account for such dynamics, it is crucial to continuously adjust the governance framework to feedback from level 3 (e.g., which quality and fairness measures should be used) and level 4 (e.g., how do street-level bureaucrats implement the ADM). If to go back to the example of the impact of the pandemic on the job market, the bureaucrats in the job agencies account for the transformed context due to the crisis in the way how they use the scores produced by the ADM (e.g., they neglect the ADM results for cases that are clearly linked to the current situation), severe adjustments to the governance framework may not be needed. However, if the ADM output is used as before, there may be the need to temporarily adjust the regulation, e.g., by issuing a new guideline about how to treat certain cases.

While this catalogue of guidelines seems to be a tall order to be implemented in decision-making processes, several empirical examples show that setting up a process that takes up these elements is not out of reach. In fact, as part of a generic development, the governance of the AMAS-ADM used in Austria to profile unemployed has already taken several steps in such a direction, for instance by involving stakeholders in the discussion about the setup of the ADM or by developing instructions on the use of the ADM ("Sozialverträglichkeitsregeln") (Holl et al., 2019). Although the implementation process may not have been ideal in other aspects (e.g., on training or on the inclusion of feedback processes from the implementation stage),[20] the example nevertheless shows that there seems to be a growing awareness about the aspects to consider when an ADM system is introduced by state authorities. Similar developments have been at work in France, where the successor platform of APB — ParcourSup — is also overseen by an ethical committee which regularly inspects problems and issues related to its use. Given these examples of governments and ministries realizing the intricacies that come with the introduction of ADM systems in certain policy areas, we are confident that the key questions developed in this paper may well set in motion a further debate about relevant aspects to be considered when state authorities ponder the question whether to buy and implement an ADM-system.

Author Contribution Not applicable.

Funding Open Access funding enabled and organized by Projekt DEAL. The research was funded by the German Ministry of Education and Research (BMBF, grant no. 16ITA203).

Data Availability Not applicable.

[20] Expert Interview September 22, 2020; Expert Interview November 3, 2020.

Declarations

Ethics Approval and Consent to Participate The research published in this article fulfills ethical standards; all interviewees consented to participate in the research.

Competing Interests The authors declare no competing interests.

References

Allhutter, D., Cech, F., Fischer, F., et al. (2020), Algorithmic profiling of job seekers in Austria: how Austerity politics are made effective. *Frontiers in Big Data* 3.

Altman, A. (2011). Discrimination. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy.* Stanford: Metaphysics Research Lab, Stanford University.

Anderson, E. (1999). What is the point of equality? *Ethics, 109*, 287–337.

Ansell, C., & Boin, A. (2019). Taming deep uncertainty: The potential of pragmatist principles for understanding and improving strategic crisis management. *Administration & Society, 51*, 1079–1112.

Arial, B., & Bland, M. (2019). Is crime rising of falling? A comparison of police-recorded crime and victimization surveys. In M. Deflem & D. M. D. Silva (Eds.), *Methods of Criminology and Criminal Justice Research* (pp. 7–32). Bingley.

Arneson, R. J. (2006). Justice after Rawls. In J. S. Dryzek & R. E. Goodin (Eds.), *The Oxford handbook of political theory* (pp. 45–64). Oxford Univ. Press.

Barfield, W. (2020). *The Cambridge Handbook of the Law of Algorithms*. Cambridge University Press.

Baumgartner, F. R., & Jones, B. D. (1991). Agenda dynamics and policy subsystems. *The Journal of Politics, 53*, 1044–1074.

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society, 20*, 1–13.

Bennett Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *Policing and Society, 28*, 806–822.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3–44.

Berner, H., & Schüll, E. (2020). Bildung nach Ma. Die Auswirkungen des AMS-Algorithmus auf Chancengerechtigkeit, Bildungszugang und Weiterbildungsförderung. *Magazin erwachsenenbildung.at. Das Fachmedium für Forschung, Praxis und Diskurs,* 40.

Blacklaws, C. (2018). Algorithms: Transparency and accountability. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 376*(2128), 20170351.

Bovens, M., Schillemans, T., & Goodin, RE. (2014). Public accountability. In M. Bovens., R. E. Goodin,.,& T. Schillemans (Eds.) *The Oxford Handbook of Public Accountability.* Oxford: Oxford University Press, https://doi.org/10.1093/oxfordhb/9780199641253.9780199641013.9780199640012.

Bowers, K. J., Johnson, S. D., & Pease, K. (2004). Prospective hot-spotting: The future of crime mapping? *The British Journal of Criminology, 44*, 641–658.

British Academy & The Royal Society. (2017). Data management and use: Governance in the 21st century. https://royalsociety.org/-/media/policy/projects/data-governance/data-management-governance.pdf

Brkan, M. (2019). Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International Journal of Law and Information Technology, 27*, 91–121.

Bullock, J. B. (2019). Artificial intelligence, discretion, and bureaucracy. *The American Review of Public Administration, 49*(7), 751–761.

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*, 220–239.

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review, 81*(5), 825–836.

Caliendo, M., Mahlstedt, R., & Mitnik, O. A. (2017). Unobservable, but unimportant? The relevance of usually unobserved variables for the evaluation of labor market policies. *Labour Economics, 46*, 14–25.

Caswell, D., Marston, G., & Larsen, J. E. (2010). Unemployed citizen or 'at risk' client? Classification systems and employment services in Denmark and Australia. *Critical Social Policy, 30*, 384–404.

Chouldechova, A., Benavides-Prado, D., Fialko, O., et al. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In A. F. Sorelle,.,& W. Christo (Eds.) *Proceedings of the 1st Conference on Fairness, Accountability and Transparency.* Proceedings of Machine Learning Research: PMLR, 134–148.

Cohen, G. A. (2009). *Why not socialism?* Princeton University Press.

Council of Europe. (2019). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. *Council of Europe, DGI, 2019*, 05.

Cour des Comptes. (2017). *Admission post-bac et accès à l'enseignement supérieur*. Cour des Comptes.

Cour des Comptes. (2018). *Un premier bilan de l'accès à l'enseignement supérieur dans le cadre de la loi orientation et réussite des étudiants*. Cour des Comptes.

Danaher, J., Hogan, MJ., Noone, C., et al. (2017). Algorithmic governance: developing a research agenda through the power of collective intelligence. *Big Data & Society* 4.

Desiere, S., Langenbucher, K., & Struyven, L. (2019). *Statistical profiling in public employment services (OECD Social, Employment and Migration Working Papers, No. 224)*. Paris: OECD Publishing.

Donia, J., & Shaw, J. A. (2021). Co-design and ethical artificial intelligence for health: An agenda for critical research and practice. *Big Data & Society, 8*(2), 20539517211065250.

Dworkin, R. (1977). Reverse discrimination. In R. Dworkin. (Ed.) *Taking Rights Seriously.* Avon: Duckworth, 223–239.

Dworkin, R. (1981). What is equality? Part 2: Equality of resources. *Philosophy and Public Affairs, 10*, 283–345.

Esser, H. (1999). Inklusion, integration und ethnische Schichtung. *Journal Für Konflikt Und Gewaltforschung, 1*, 5–34.

Commission, E. (2020). *On artificial intelligence - a European approach to excellence and trust COM/2020/65 final*. WHITE PAPER.

Franke, U. (2021). Rawls's original position and algorithmic fairness. *Philosophy & Technology, 34*(4), 1803–1817.

Fullinwider, R. (2018). Affirmative action. In E. N. Zalta (Ed.) *Stanford Encyclopedia of Philosophy.* Stanford: Metaphysics Research Lab, Stanford University.

Gajduschek, G. (2003). Bureaucracy: Is it efficient? Is it not? Is that the question?: Uncertainty reduction: An ignored element of bureaucratic rationality. *Administration & Society, 34*, 700–723.

Gamper, J., Kernbeiß, G., & Wagner-Pinter, M. (2020.) *Das Assistenzsystem AMAS. Zweck, Grundlagen, Anwendung.* Wien: Synthesisforschung GmbH.

Gillingham, P. (2019). Can predictive algorithms assist decision-making in social work with children and families? *Child Abuse Review, 28*, 114–126.

Granovetter, M. S. (1995). *Getting a job. A study of contacts and careers*. The University of Chicago Press.

Gritsenko, D., & Wood, M. (2020a). Algorithmic governance: a modes of governance approach. *Regulation & Governance* first view.

Gritsenko, D., & Wood, M. (2020b.) Algorithmic governance: a modes of governance approach. *Regulation & Governance*, first view.

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*, 205–211. https://doi.org/10.1136/medethics-2019-105586

Habermas, J. (2021). Überlegungen und Hypothesen zu einem neuen Strukturwandel der politischen Öffentlichkeit, In M. Seeliger & S. Sevignani (Eds.) *Ein neuer Strukturwandel der Öffentlichkeit? Sonderband Leviathan, 37*. Baden-Baden: Nomos Verlagsgesellschaft.

Haeri, M. A., & Zweig, K. A. (2020). The crucial role of sensitive attributes in fair classification. *IEEE Symposium Series on Computational Intelligence (SSCI), 2020*, 2993–3002.

Harkens, A., Achtziger, A., Felfeli, J., et al. (2020). The rise of AI-based decision-making tools in criminal justice: Implications for judicial integrity. *Commonwealth Judicial Journal, 25*, 18–26.

Harris, H. M., Goss, J. G., & Gumbs, A. (2019). *Pretrial risk assessment in California*. Public Policy Institute of California.

Hartmann, K., & Wenzelburger, G. (2020). Uncertainty, risk and the use of algorithms in policy decisions A case study on criminal justice in the US. *Policy Sciences* forthcoming.

Heidari, H., Loi, M., Gummadi, K.P., and Krause, A. (2019). A moral framework for understanding fair ML through economic models of equality of opportunity. *Proceedings of the Conference on Fairness, Accountability, and Transparency.* Atlanta, GA, USA, Association for Computing Machinery.

Heinz, W. R. (1999). *From education to work: Cross national perspectives*. Cambridge University Press.

Holl, J., Kernbeiß, G., & Wagner-Pinter, M. (2018). *Das AMS-Arbeitsmarktchancen-Modell*. Synthesisforschung GmbH.

Holl, J., Kernbeiß, G., & Wagner-Pinter, M. (2019). *Personenbezogene Wahrscheinlichkeitsaussagen (»Algorithmen«) Stichworte zur Sozialverträglichkeit*. Synthesis Forschunggesellschaft GmbH.

Holton, R., &Boyd R. (2020). 'Where are the people? What are they doing? Why are they doing it?'(Mindell) Situating artificial intelligence within a socio-technical framework. *Journal of Sociology* first view, 1440783319873046.

House of Lords (2018). *AI in the UK: ready, willing and able?* HL Paper 100. Select Committee on Artificial Intelligence, Report of Session 2017–19.

Hudson, L. (2017) *Technology is biased too. How do we fix it?* FiveThirtyEight, 2017. https://fivethirtyeight.com/features/technology-is-biased-too-how-do-we-fix-it/

Jörke, D. (2013). Re-Demokratisierung der Postdemokratie durch alternative Beteiligungsverfahren? *Politische Vierteljahresschrift, 54*, 485–505.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth A. (2017). *Rawlsian fairness for machine learning*. arXiv:1610.09559v2

Jugov, T., & Ypi, L. (2019). Structural injustice, epistemic opacity, and the responsibilities of the oppressed. *Journal of Social Philosophy, 50*, 7–27.

Katzenbach, C,. & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review,* 8.

Killias, M., Aebi, MF., Aubusson de Cavarlay, B., et al. (2010.) *European Sourcebook of Crime and Criminal Justice Statistics – 2010.* Den Haag: WODC.

Kim, S., Andersen, K.N. & Lee, J. (2021). Platform government in the era of smart technology. *Public Administration Review*, online first.

Klingel, A., Krafft, TD.,& Zweig, KA. (2020). Mögliche Best Practice-Ansäztebeim Einsatz eines algorithmischen Entscheidungsunterstützungssystems des AMAS-Algorithmus. In M. Hengstschläger, & Rat für Forschung und Entwicklung (Eds.) *Digitaler Wandel und Ethik.* Salzburg und München: EcoWinVerlag, 190–215.

König, PD.,& Krafft, TD. (2020). Evaluating the evidence in algorithmic evidence-based decision-making: the case of US pretrial risk assessment tools. *Current Issues in Criminal Justice* forthcoming.

König, PD., & Wenzelburger, G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly,* 37, 101489.

König, P.D., & Wenzelburger, G. (2021). The legitimacy gap of algorithmic decision-making in the public sector: Why it arises and how to address it. *Technology in Society*, 67, 101688.

Krafft, TD., Zweig, KA., & König, PD. (2020). How to regulate algorithmic decision-making: a framework of regulatory requirements for different applications. *Regulation & Governance* first view.

Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. *AI and Ethics, 1*(4), 529–544.

Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society, 16*, 252–260. https://doi.org/10.1108/JICES-06-2018-0056

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611–627.

Lodge, M.,& Mennicken, A. (2019). Reflecting on public service regulation by algorithm, In K. Yeung, & M. Lodge (Eds.) *Algorithmic Regulation*. Oxford: Oxford University Press.

Lopez, P. (2020). Reinforcing intersectional inequality via the AMS algorithm in Austria. *Proceedings of the STS Conference.* Graz:https://openlib.tugraz.at/download.php?id=5e29a88e0e34f&location=browse.

Martin, K. (2018). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 160*, 835–850. https://doi.org/10.1007/s10551-018-3921-3

Martini, M., Botta, J., Nink, D., et al. (2020). *Automatisch erlaubt? Fünf Anwendungsfälle algorithmischer Systeme auf dem juristischen Prüfstand.* Gütersloh: Bertelsmann Stiftung.

Martini, M. (2019). *Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz*. Springer.

Matthew, J., Kearns, MJ., Morgenstern, J., et al. (2016). Rawlsian fairness for machine learning. *CoRR* abs/1610.09559.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitionS. *Annual Review of Statistics and Its Application, 8*(1), 141–163.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 2053951716679679.

Nagel, T. (1973). Equal treatment and compensatory discrimination. *Philosophy & Public Affairs, 2*, 348–363.

Noriega-Campero, A., Bakker, M., Garcia-Bulle, B., & Pentland, A. (2019). *Active fairness in algorithmic decision making*. AIES '19, January 27–28, 2019, Honolulu, HI, USA. https://doi.org/10.1145/3306618.3314277

Oswald, M., Grace, J., Urwin, S., et al. (2018). Algorithmic risk assessment policing models: Lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law, 27*, 223–250.

Pettit, P. (1997). *Republicanism. A theory of freedom and government*. Oxford University Press.

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Rawls, J. (1999). *A theory of justice* (Revised). Belknap Press.

Rawls, J. (2005). *Political liberalism* (expanded). Columbia University Press.

Robertson, S., Nguyen, T,.& Salehi, N. (2021). Modeling assumptions clash with the real world: transparency, equity, and community challenges for student assignment algorithms. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Yokohama, Japan, Association for Computing Machinery.

Schwartz, J., & Vega, A. (2017). Sources of crime data. In B. Teasdale & M. S. Bradley (Eds.), *Preventing Crime and Violence* (pp. 155–167). Springer International Publishing.

Segal, S., Adi, Y., Pinkas, B., Baum, C., Ganesh, C., & Keshet, J. (2021). Fairness in the eyes of the data: certifying machine-learning models, In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*: Association for Computing Machinery.

Shah, K., Gupta, P., Deshpande, A., & Bhattacharyya, C. (2021). Rawlsian fair adaptation of deep learning classifiers, In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*: Association for Computing Machinery.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*, 337–356.

Sirsch, J. (2013). Die Regulierung von Hassrede in Liberalen Demokratien. InJ. Meibauer (Ed.) *Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion.* Gießen: Gießener Elektronische Bibliothek, 165–194.

Thomson, J. J. (1973). Preferential hiring. *Philosophy & Public Affairs, 2*, 364–384.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: Key problems and solutions. *AI & SOCIETY, 37*(1), 215–230.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Uher, J. (2019). Data generation methods across the empirical sciences: Differences in the study phenomena's accessibility and the processes of data encoding. *Quality & Quantity, 53*, 221–246.

UK Statistics Authority. (2014). *Statistics on crime in England and Wales. Assessment Report 268.* London: UK Statistics Autority.

Ulbricht, L., & Yeung, K. (2022). Algorithmic regulation: A maturing concept for investigating regulation of and through algorithms. *Regulation & Governance, 16*(1), 3–22.

van der Voort, H. G., Klievink, A. J., Arnaboldi, M., et al. (2019). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly, 36*, 27–38.

Veale, M., van Kleek, M., & Binns, R. (2018). *Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making*. *CHI 2018.* Montréal, Canada.

Vedder, A., & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers & Technology, 31*, 206–224.

Vis, B. (2018). Heuristics and political elites' judgment and decision making. *Political Studies Review* forthcoming.

Waldron, J. (2012). *The harm in hate speech (The Oliver Wendell Holmes Lectures, 2009)*. Harvard University Press.

Wenzelburger, G., & Hartmann, K. (2021). Policy formation, termination and the multiple streams framework: the case of introducing and abolishing automated university admission in France. *Policy Studies*, 1–21.

P-H Wong 2020 Democratizing Algorithmic Fairness. Philosophy & Technology 33 2 225 244

Yeung, K. (2018a). Algorithmic regulation: A critical interrogation. *Regulation & Governance, 12*, 505–523.

Yeung, K., & Lodge, M. (2019). *Algorithmic regulation*. Oxford University Press.

Yeung, K. (2018b). Algorithmic regulation: A critical interrogation. *Regulation & Governance, 12*(4), 505–523.

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance, 2*(4), 301–313.

Završnik, A. (2019). Algorithmic justice: algorithms and big data in criminal justice settings. *European Journal of Criminology* online first.

Zouridis, S., van Eck, M., & Bovens, M. (2020). Automated discretion. In T. Evans, & P. Hupe (Eds.) *Discretion and the Quest for Controlled Freedom*. Cham: Springer International Publishing.

Authors and Affiliations

Maryam Amir Haeri[1] · Kathrin Hartmann[2] · Jürgen Sirsch[3] · Georg Wenzelburger[2] · Katharina A. Zweig[2]

Maryam Amir Haeri
m.amirhaeri@utwente.nl

Kathrin Hartmann
kathrin.hartmann@sowi.uni-kl.de

Jürgen Sirsch
sirsch@uni-mainz.de

Katharina A. Zweig
zweig@cs.uni-kl.de

[1] University of Twente, Enschede, Netherlands

[2] Department of Social Sciences, TU Kaiserslautern, Postbox 3049, 67653 Kaiserslautern, Germany

[3] Johannes Gutenberg-University, Mainz, Germany