Krafft, Tobias D.; Zweig, Katharina A.; König, Pascal D.

How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications

Regulation & Governance

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Krafft, Tobias D.; Zweig, Katharina A.; König, Pascal D. (2022) : How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications, Regulation & Governance, ISSN 1748-5991, John Wiley & Sons Australia, Ltd, Melbourne, Vol. 16, Iss. 1, pp. 119-136, https://doi.org/10.1111/rego.12369

Regulation & Governance

Regulation & Governance (2022) 16, 119–136 doi:10.1111/rego.12369

How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications

Tobias D. Krafft, Katharina A. Zweig

Algorithm Accountability Lab at the Department of Computer Science, University of Kaiserslautern, Kaiserslautern, Germany

Pascal D. König

Department of Social Sciences, University of Kaiserslautern, Kaiserslautern, Germany

Abstract

Algorithmic decision-making (ADM) systems have come to support, pre-empt or substitute for human decisions in manifold areas, with potentially significant impacts on individuals' lives. Achieving transparency and accountability has been formulated as a general goal regarding the use of these systems. However, concrete applications differ widely in the degree of risk and the accountability problems they entail for data subjects. The present paper addresses this variation and presents a framework that differentiates regulatory requirements for a range of ADM system uses. It draws on agency theory to conceptualize accountability challenges from the point of view of data subjects with the purpose to systematize instruments for safeguarding algorithmic accountability. The paper furthermore shows how such instruments can be matched to applications of ADM based on a risk matrix. The resulting comprehensive framework can guide the evaluation of ADM systems and the choice of suitable regulatory provisions.

Keywords: accountability, agency theory, algorithmic decision-making, risk matrix, transparency.

1. Introduction

Individuals in information societies are increasingly subject to the scoring and classification performed by algorithmic decision-making (ADM) systems (Saurwein *et al.* 2015). ADM systems have found their way into such diverse fields as online advertisement, medical diagnosis, credit lending, job applicant selection, and risk assessments in criminal justice (Fry 2018). In all of these cases, ADM systems solve specialized cognitive tasks in order to ultimately inform some sort of intervention or treatment broadly understood.

As these systems increasingly inform or substitute for human decision-making, possibly with profound consequences for individuals' welfare or through extensively intervening into social relations (Brauneis & Goodman 2018; Just & Latzer 2017; Yeung 2017b; Ulbricht & Yeung 2022) algorithmic accountability has become an intensely researched topic. An important point of reference in this regard is the General Data Protection Regulation (GDPR) of the European Union. While it provides a general framework that can be used for regulating automated decision-making, it does not comprehensively and clearly state how instruments for safeguarding transparency and accountability can be applied to different applications of ADM (Brkan 2019; Bygrave 2019). As a study by the European Parliamentary Research Service has indicated, the GDPR is "not likely to be sufficient" (Koene *et al.* 2019, p. III) to adequately safeguard accountability of ADM systems.

A rich and quickly growing scholarly literature spanning different disciplines deals with the challenges of algorithmic accountability. Contributions on algorithm ethics (Mittelstadt *et al.* 2016; Binns 2017; Ananny & Crawford 2018; Lanzing 2018) and from the fields of law and regulation (Hildebrandt 2008; Koops 2013;

Correspondence: Tobias D. Krafft, Department of Computer Science, University of Kaiserslautern, PO-Box 3049, 67663 Kaiserslautern, Germany. Email: krafft@cs.uni-kl.de

Conflict of interest: The authors declare that they have no conflict of interest.

Accepted for publication 5 October 2020.

Hildebrandt 2016; Brauneis & Goodman 2018; Yeung 2017a) point to core ethical principles and procedural criteria which can make sure that ADM systems do not cause harm to or violate the rights of those affected by their decisions. Other, more technically informed work furthermore points to various ways in which ADM systems can be made transparent and subjected to control (Diakopoulos 2014; Kroll *et al.* 2017; Guidotti *et al.* 2018; Lepri *et al.* 2018; Bryson & Theodorou 2019; Sokol & Flach 2020; Wieringa 2020).

These strands in the literature point to principles or concrete instruments as ways to safeguard algorithmic accountability in general. They are, however, not concerned with systematically distinguishing between different uses of ADM systems and how these can be accommodated with adequate regulatory provisions. Yet given the various purposes for which ADM systems can be used, effective regulation for ensuring transparency and accountability could take very different forms – ranging from minimal or no requirements to very rigorous provisions. As Nemitz (2018, p. 8) has remarked, there is a need to turn to the question which challenges of algorithmic accountability need to be addressed by rules and enforceable requirements and which can do without. Some contributions looking at specific fields of applications of ADM have furthermore cautioned that one size does not fit all when it comes to governing the risks of algorithmic systems (Saurwein *et al.* 2015; Krafft & Zweig 2019; van Drunen *et al.* 2019).

This work suggests that a differentiated regulatory approach to ADM systems is important because not all uses are equally problematic. These systems can be embedded in very different settings and vary widely in terms of their purposes as well as the decision consequences and the risks involved. This means that regulatory provisions need to accommodate different applications of ADM systems. Otherwise, regulation might be too restrictive and hamper initiative and innovation in some areas or be too weak in others. Presuming that it is generally desirable for society that effective control and regulation occur at the smallest cost possible (Jensen & Meckling 1976), the intensity of regulation and thus the modes and instruments for achieving accountability in the use of ADM systems needs to be adapted to different uses.

Against this backdrop, the present paper contributes to existing research by formulating a differentiated framework for governing ADM systems. The framework systematizes ways of safeguarding accountability that serve to match regulatory provisions to a range of situations in which ADM systems are used. We draw on agency theory as a theoretical lens for conceptualizing accountability challenges and how these can be addressed. Specifically, we subsume instruments for safeguarding algorithmic transparency and control under different accountability mechanisms. Starting from this conceptual footing, we then go on to show how those instruments can be matched to different applications of ADM based on a risk matrix. The result is a comprehensive framework that points to important principles and criteria for a differentiated regulation of a wide range of ADM systems.

2. ADM systems and issues of transparency and accountability

2.1. ADM systems and their variable impacts on individuals and society

The general idea behind ADM systems is to use information about entities and their behaviors[1] in order to assign them a single numeric value by means of clearly defined instructions, that is, through an algorithm. This assigned value then informs some decision or intervention that is either fully automated or occurs with a human in the loop. In some contexts, the affected entities could be networked machines whose operations are coordinated (e.g. in a so-called smart factory). Where ADM systems are adopted in a social context, those entities are usually individuals. Given the potentially more far-reaching societal impact and ethical consequences in this latter setting, we will focus on how ADM systems serve to make decisions for and/or about individuals. This can take the form of scoring, for instance, where individuals are attributed a number that might express a risk, such as credit default; or the goal might be a classification, such that the resulting value corresponds to a specific category, for example a discrete class of consumer preferences.

In the case of so-called expert systems, the rules for arriving at such value attributions have been formulated and formalized as detailed decision trees by humans. In contrast, machine learning inductively generates decision rules based on patterns identified in data, for example about individuals and their behaviors (Watt *et al.* 2020). Machine learning used in ADM systems commonly requires a so-called ground truth to acquire decision rules. Taking the example of consumer behavior prediction for targeted advertising, this would mean that there exists information about whether recommended products have indeed been bought. Based on this information, the ADM system can learn which recommendations in combination with which circumstances are most likely to be

effective. Learning in this case means to identify those features of individuals that correlate most strongly with actual buying behavior (which is provided to the system in the form of the ground truth). In sum, a learning ADM system builds a *statistical model* that is supposed to represent a specific part of reality. Assuming that previous behaviors are indicative of future behaviors, this statistical model serves to predict the probability of buying decisions depending on the combinations of features that describe potential buyers.

Hence, ADM systems that are based on machine learning, strictly speaking, consist of two kinds of algorithms. The first algorithm serves to infer decision rules from data whereas the second algorithm merely uses these decision rules to score or classify cases. The core of the ADM system is therefore the first of these two algorithms – the learning method – and the decision rules generated from it, which can take very different forms and be quite complex. The scoring or classification algorithm, in contrast, is usually rather simple as it merely applies the trained statistical model (Zweig *et al.* 2018).

ADM systems are generally designed to deal with some specialized cognitive tasks. Even where they are technically similar, however, the consequences of the decisions and interventions informed or determined by ADM systems may differ considerably depending on the concrete setting in which these systems are applied. Accordingly, the kind of risks for those about and/or for whom decisions are made can vary widely – as it is the case with human decision-making. ADM systems might be part of a service to consumers with minimal to no risk. Other applications may affect individuals' welfare and life chances, for example where it is used to select the unemployed for job qualification offers, to decide about credit lending or to inform the choice of medical treatments.

Given the breadth of possible applications of ADM systems, they can produce all kinds of risks that are well known from research on consumer behavior, such as functional, physical, social, and financial risks (Schiffman *et al.* 2012, p. 197). In a broader perspective, one could also include such diverse risks as violation of intellectual property rights, harm to privacy, and abuse of market power (Saurwein *et al.* 2015, p. 37). However, as our focus is on algorithmic accountability, the risks that we are concerned with are those that are directly tied to decision-making and the consequences resulting from it.[2]

Among these risks, discrimination and adverse effects on individuals' autonomy have received particular attention in the literature, arguably because they can take rather subtle and partly novel forms with the use of ADM systems. First, ADM systems may exhibit biases in the form of discrimination based on sensitive features, such as gender or ethnic group. For instance, the algorithmic recommendation of job qualification measures for unemployed persons might, ceteris paribus (i.e. with other characteristics being the same for different persons), give lower scores to members of a certain ethnic group. Such a bias can be acquired from processed training data in which patterns of unfair discrimination are already represented (Barocas & Selbst 2016, pp. 674–675; Mittelstadt *et al.* 2016, pp. 8–9). Dealing with such a learned bias is a value-laden affair because there are different ways in which fairness can be mathematically formalized and these are partly incompatible with each other (for an overview Berk *et al.* 2018, pp. 17–23). Operationalizing fairness into a machine hence requires a clear understanding of what fairness and absence of discrimination means in the first place.

Secondly, it has been noted that ADM systems may lead to heteronomy as they can structure social relations and individual behavior according to certain objectives that are not known to those affected. In the context of social network sites, for instance, previous research has pointed out that ADM systems serve to shape the information environments and the choice situations of many individuals in a personalized fashion, thereby making some behaviors more likely than others (Mittelstadt *et al.* 2016; Just & Latzer 2017; Yeung 2017a; Yeung 2017b; Eyert *et al.* 2022). This subtle influence is made possible through fine-grained data, the extraction of behavioral patterns, and predictions of individuals' likelihood of future behaviors. Affected individuals may well be happy with the provided content and the overall experience of a service. However, as Lanzing (2018, p. 11) has argued, such practices may be an unjustified interference with one's decisional autonomy, as a person "can no longer be certain whether they are acting based on their own reasons, reasons they selected themselves and identify with.".

All in all, while implementations of ADM systems may have similar technical properties, their effects on individuals and on society, as well as the risks they entail may differ widely depending on the setting and the way in which they are adopted. A regulatory approach to the use of ADM systems that accommodates these considerations therefore has to be both broad in terms of covering a wide range of applications, and differentiated, that is, through taking into account relevant differences concerning the risk for affected individuals.

2.2. On the need for a differentiated regulatory approach

A major point of reference with regard to the regulation of algorithmic data processing is the European GDPR, which contains various provisions for safeguarding transparency and accountability in the processing of personal data, including by ADM systems. However, the GDPR is primarily designed to govern the protection of personal data and not to regulate ADM systems. Indeed, scholarly debate about the GDPR has pointed to important limitations in addressing challenges of achieving fair, transparent, and accountable ADM systems. First, the range of applications of ADM that falls under relevant GDPR provisions is vaguely defined. Article 22 refers to a right not to be subjected to solely automated decision-making, including profiling, that has legal consequences or similarly significant effects. Its application is therefore restricted to ADM systems with certain consequences, and it has been noted, it may leave a loophole as a marginal human intervention may prevent decision-making from counting as "solely" automated (Edwards & Veale 2017; Brkan 2019; Bygrave 2019).

Secondly, the transparency and disclosure provisions in the GDPR are very broad. Articles 13 to 15 guarantee the individual data subject a right to obtain "meaningful" information about the logic of automated decision-making and its consequences. Equally, the safeguards that are to be taken where automated decision-making is applied, for example when based on explicit consent, remain rather vague in the GDPR (Brkan 2019). It states as general objectives that these safeguards are supposed to protect the data subject's rights, freedoms, and legitimate interests. Scholarly debate about these safeguards has dealt with the question whether they imply a right to an explanation and discussed the usefulness of explanation without intelligibility of ADM systems (for an overview, see Malgieri 2019). Yet, even if one presumes a right to explanation and legibility of automated decision-making this could take very different forms in practice. Overall, the GDPR is a flexible normative and regulatory basis that does, however, not offer firm guidance on which instruments for transparency and accountability can be employed to accommodate different kinds of application of ADM (Bygrave 2019, p. 260; Gellert 2022).

Given the range of ways in which ADM systems can be implemented, a differentiated regulatory approach seems necessary. Regarding the right to an explanation, van Drunen *et al.* (2019) argue that a single, uniform way of applying that right would hardly do justice to the different uses of automated data processing. Rather, the way in which such an explanation is realized (e.g. the kind of information disclosed) would need to be adapted to the context in which an ADM system operates. As van Drunen *et al.* (2019) are concerned with news and content filtering, they focus on different forms that a right to an explanation may take in that context. However, such a right is itself only one way of achieving algorithmic accountability and not the most appropriate for all contexts. It may enable a person to gain an understanding of how decisions are made, but this is not be helpful with all uses of ADM systems. Specifically, getting an explanation does not mean that a person can trust that due diligence has been followed in the design of an ADM system (on this see Bryson & Theodorou 2019, p. 317), for example, for best reducing undesirable biases or unfair outcomes that are hard to detect, especially for an individual. Transparency about these aspects will require a different form of disclosure.

A broader perspective on the regulation of ADM systems has been taken by Saurwein *et al.* (2015). While they narrowly speak of "algorithmic selection," they consider a broad range of purposes of ADM systems and point to various risks. The authors discuss possible regulatory means to deal with these risks and argue that one size does not fit all in the governance of risks in algorithmic selection. Rather, different actors – individuals, industry, and the state – will have to be involved in variable constellations. While the approach by Saurwein and colleagues considers a broad range of potential risks, they compile a set of risks that goes well beyond the decision-making itself (they include, for instance, market abuse or violation of intellectual property rights), and their focus is on the responsibility of different actors for regulating these systems depending on the risks involved.

In the following, we adopt a different perspective that focuses on risks which are tied to the decision-making as such. Starting from a set of accountability challenges that arise from the use of learning ADM systems, we systematize suitable provisions for addressing these challenges. In doing so, we subsume various concrete instruments for achieving transparency and accountability of algorithms that have been proposed in a more technical literature (e.g. Diakopoulos 2014; Kroll *et al.* 2017; Lepri *et al.* 2018; Bryson & Theodorou 2019) under distinctive accountability mechanisms.

3. Accountability challenges in the adoption of ADM systems

The principal-agent model is a key conceptual paradigm in the regulation literature used to study control problems and mechanisms in accountability relationships (see e.g. Lodge 2004). In that model, a principal relies on an agent to fulfill a task and act in the principal's interest. While the utility for the principal depends on the performance of the agent, the agent may reduce that utility through pursuing her own interest. The model has been developed to identify challenges for a principal in the delegation of decision-making and to examine the ways in which the agent's behavior can be aligned with the interest of the principal (Pratt & Zeckhauser 1991; Lane 2007).

Classical accounts of principal-agent relations from economics (Hölmstrom 1979) and political science (Weingast & Moran 1983; McCubbins & Schwartz 1984) commonly take organizations as the principals delegating tasks to an agent, for example, businesses contracting managers or a parliament overseeing the bureaucracy. The use of ADM systems could also be analyzed in these terms as the state and businesses employ them for various purposes and to better achieve their objectives, partly through a service provided by third parties. For instance, the state might use predictive policing or a business may want to harness ADM systems for placing targeted online advertisements. However, the delegation of decision-making in these cases concerns the questions how the use of an ADM system can best be oriented toward the goals of these organizations, such as increasing public safety or maximizing profit.

In the following, we adopt a different perspective and start from accountability problems that arise from the point of view of individuals as data subjects. This perspective more directly brings into view ethically relevant consequences of ADM systems for individuals (see Mittelstadt *et al.* 2016) and it is also congruent with the outlook of the literature on fair, accountable, and transparent algorithms (e.g. Pasquale 2015; Lepri *et al.* 2018). The accountability relationship we focus on is, therefore, similar to the principal-agent relation found for clients or consumers contracting professional services (Robinson *et al.* 2010, p. 19; Heremans 2012, p. 31).

Indeed, various ADM systems are part of a service that data subjects use (by agreeing to terms and conditions) and that they expect to operate in their interest. However, issues of algorithmic accountability range further as an accountability relationship may also exist where no formal contracting and delegation take place. Such a constellation can be accommodated with a broader conceptualization of principal-agent relations, which presumes that delegation does not have to be formal but can also be informal (Hill & Jones 1992, p. 134; Kerwer 2005, p. 457). Although those affected by the decision-making do not directly delegate to an agent, they can nevertheless be seen as "external principals" (Kerwer 2005, p. 458) who have legitimate expectations about the behavior of the agent – and thus have grounds for holding the agent to account by some standard (Pratt & Zeckhauser 1991).

Based on this broader understanding of agency problems, the accountability relationship that we are concerned with exists where data subjects provide inputs, including in the form of personal data, that lead to outputs from an ADM system; and these outputs inform decisions for and/or about the data subject in a way that concerns the subject's legitimate interests.

This perspective includes settings in which, first, ADM outputs are provided *for* data subjects, in which they are service users consenting to terms and conditions. They might be users of a social media news feeds that performs a sorting task for them, and one that is personalized based on a profiling of individuals. Other examples are search engines, voice-controlled personal assistants, dating apps, or medical diagnosis tools that all perform a task for data subjects.[3] In all these cases, individuals request a service, they provide inputs and expect a processing of data that yields outputs which are produced for them and align with their goals.

Other ADM systems, second, make explicit scoring and classification decisions *about* data subjects that are, however, not solicited by them as a service and are in that sense not made *for* them. This is the case, for example with microtargeted online advertisements and personalized pricing, but also tax fraud detection tools, credit default risk assessments, or recruitment tools for selecting applicants. Clearly, affected subjects in these latter settings cannot expect to get positive decisions, but they do have legitimate expectations, for instance, with regard to not being discriminated against based on sensitive features. Even without formal delegation or contracting, data subjects are still external principals because an ADM system produces decisions about the data subject that affect her welfare or personal rights.

In all these constellations, there is an accountability relationship to the extent that affected subjects have a legitimate interest in ADM systems respecting their preferences, autonomy, or personal rights. Affected subjects can expect the ADM system to perform in a certain way and have grounds to hold the decision-making to account by certain standards, even without formal delegation. This accountability relationship along the lines of a principal-agent relation is depicted in Figure 1.

It should be noted that the ADM system alone is not, strictly speaking, the agent in the accountability relationships. Rather, it must be seen in conjunction with those who commission, design and implement the ADM system, for example as part of a service. For the sake of simplicity, we will, however, speak of ADM systems as agents in the following. There are several reasons why this still captures essential aspects of accountability relations where ADM systems are employed. First, it has repeatedly been pointed out that ADM systems are never neutral but necessarily incorporate certain assumptions, goals, and values (e.g. Hildebrandt 2016; Mittelstadt *et al.* 2016; Just & Latzer 2017; Yeung 2017b). This means that these systems themselves realize certain objectives, usually derived from those actors who implement and operate them. Secondly, ADM systems can also show a certain degree of autonomy as they can acquire decisions rules and biases that were not anticipated, and may therefore even be seen as agents themselves (Mittelstadt *et al.* 2016, p. 3). Finally, realizing accountability regarding the use of ADM systems also has to accommodate certain technical features of the systems themselves.

In sum, an ADM system can be understood as an agent that realizes certain values and objectives; and these may go against the legitimate expectations and interests of the data subjects for and/or about whom decisions are produced. At the same time, these data subjects may not have the information needed to assess the ADM system and its performance. Altogether, this constellation is characterized by the same fundamental challenges as in achieving congruence between the behavior of the agent and the interests of the principal: diverging interests and an information advantage of the agent over the principal (Sappington 1991; Miller 2005). As an agent makes use of her discretion and exploits information asymmetries, the principal experiences *agency loss*. This notion of agency loss refers to the difference between the agent's actual behavior and its behavior if it conformed to the principal's legitimate expectations (Pratt & Zeckhauser 1991; Lupia 2003). The information asymmetries that a principal generally faces and that can lead to agency loss take different forms. They concern not only the agent's motives and qualities, but also the knowledge/information the agent possesses and the agent's – possibly hidden – actions (Saam 2007, p. 827).

These general kinds of information asymmetries also apply to the relation between data subjects and ADM systems, as Figure 1 illustrates. First, relevant qualities of the ADM system, including the objectives (the agent's "intentions") that it realizes, may be unknown to the affected data subjects. They may, therefore, be unaware that the system or the service on which it is based realizes goals that diverge from the data subjects' interests. This could be the case because the ADM system does not perform well, for example, makes many wrong classifications (a hidden quality comparable to incompetence). However, even if an ADM system achieves an expected quality it may still embody certain objectives and values that go against the expectations and interests of affected data subjects as (external) principals. This could be due to deliberate choices in the design of an ADM system, but it may also be an unintended result: If the processed data contains unknown and undesired biases, for example, in the form of a discrimination based on sensitive features, these may be learned by the ADM system (Barocas & Selbst 2016; Lepri *et al.* 2018). There is thus the possibility that an ADM system embodies values and assumptions that were not intended as part of the design.

A second challenge that the principal faces is hidden knowledge/information. The agent may enjoy an information advantage that can be exploited to take actions which serve her own interest. Even if the principal can observe agent behavior, she may not be able to evaluate the quality of the agent's decisions. This problem can again also occur with ADM systems, affected data subjects may not know the parameters according to which decisions are made. The ADM system might furthermore process information about data subjects (e.g. preferences and behavior patterns) that can be used to alter subjects' behaviors by purposefully structuring decision situations in ways that they are not aware of (Yeung 2017a).

A third major challenge for a principal are hidden actions by the agent, that is, actions that can only be revealed at high costs, which gives the agent some discretionary space to change the amount of effort put in fulfilling some task (Hölmstrom & Milgrom 1987). Hidden action is an acute challenge with ADM systems. From the perspective of data subjects, these systems commonly operate unobtrusively in the background while the complexity,

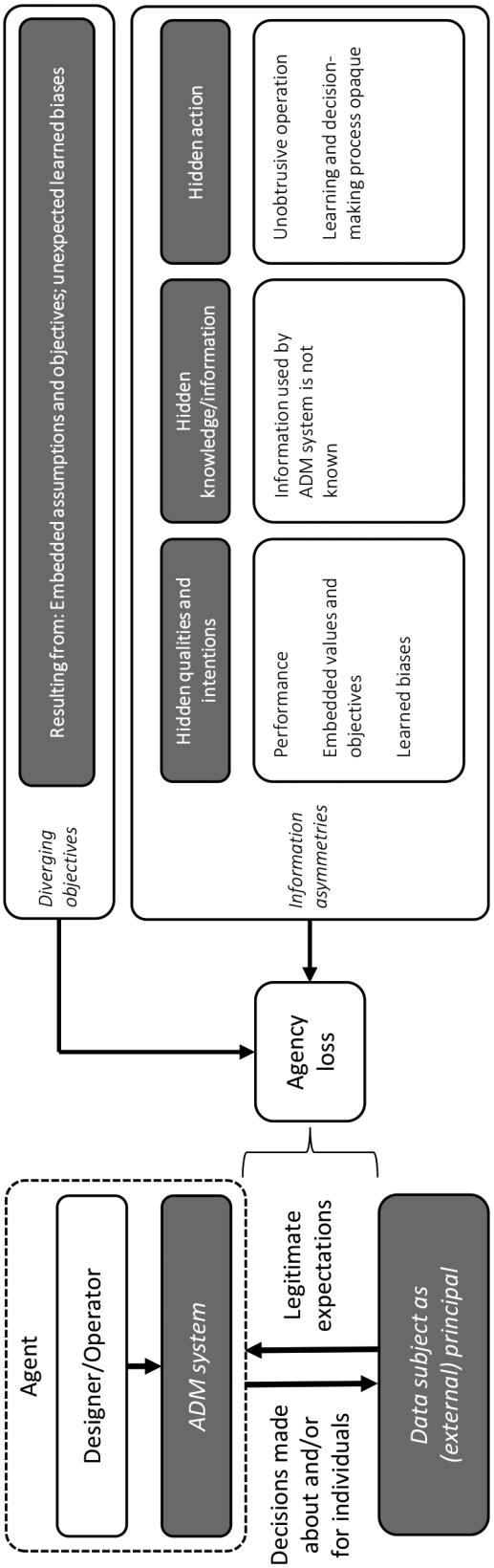

Figure 1 Information asymmetries in principal-agent relations applied to algorithmic decision-making.

speed, and scale of its operations make it hard to observe decisions made by the system and the learning processes that it undergoes (Mittelstadt *et al.* 2016, p. 6). This too introduces scope for an undesired behavior of the system.

In sum, the information asymmetries that lead to agency loss in accountability relationships in general can straightforwardly be mapped onto data subjects facing agency problems vis-à-vis ADM systems, as Figure 1 underscores. Against this backdrop, we use the principal-agent model as a conceptual paradigm to identify which accountability mechanisms and corresponding concrete instruments may be needed to mitigate challenges of algorithmic accountability.

This is not to say, however, that the data subjects themselves are generally able to address and remedy the accountability deficits. Indeed, as the mechanisms and instruments for safeguarding algorithmic accountability described below make apparent, placing the burden on the individual data subject is hardly an option in most cases. This will often hardly be possible, not only because of the time and expertise needed and the systemic nature of some accountability problems, but also because intellectual property rights and the possibility of gaming ADM systems are valid reasons for not publicly disclosing certain information about it (Pasquale 2015; Veale *et al.* 2018; Zweig *et al.* 2018). Institutional solutions will thus be needed to safeguard the accountability of ADM systems. This is not unlike the setting in which consumers are principals expecting a service from an agent, but where regulators act on behalf of consumers to mitigate agency loss (see e.g. Weiss 1995, pp. 69–72; Sherman 2011).

In light of the variegated uses of ADM systems and their potential complexity, the state and its regulatory bodies will also likely have to rely on what has been called "regulatory intermediaries" (Abbott *et al.* 2017, p. 19), actors that are – formally or informally – acting together with a regulator to shape the behavior of a given target. Indeed, it has been argued with a view to ADM systems, that a range of different actors will have to be involved in efforts to oversee and control the use of such systems in certain areas (Saurwein *et al.* 2015). Regardless of which concrete actors are ultimately tasked to safeguard accountability of ADM systems, the framework developed below points to where and how regulatory action can intervene. Its aim is to identify and systematize suitable regulatory means for dealing with accountability problems from the point of view of affected data subjects.

4. How to safeguard accountability of ADM system behavior

Based on the principal-agent model, one can identify several ways in which agency loss can be reduced. These differ with regard to whether direct incentives are created to induce desired agent behavior or whether certain procedures are employed to reduce information asymmetries, such as reporting requirements and monitoring mechanisms (Nalebuff & Stiglitz 1983; Lupia 2003; Miller 2005). Depending on which means are used to keep the agent in check in general, one can distinguish between different methods for safeguarding the agent's accountability. Applied to the use of ADM systems, we distinguish between four such mechanisms that are summarized in Figure 2 and will be detailed in the following.

Accountability is constituted by transparency and answerability – having to provide information and justifications – but also requires enforcement, that is, the possibility of sanctioning, without which any transparency is hardly of any use (Warren 2014). As a minimal form that does not rely on special transparency and answerability provisions, accountability – and desired agency behavior – can be achieved through inducing incentives which result from choice and competition in the market. This *market accountability* (Hirschmann 1970) is realized where there is a substantive choice between options and it is easy to switch between agents. As long as a principal can trust that the directly perceptible outcomes of agent behavior are sufficient for staying with or switching the agent, this simple way of market-based sanctioning through selecting and switching between agents can be effective in deterring unwanted agent behavior. Under conditions of effective competition, the agent's stakes of not performing in the interest of a principal are increased and form an important source of incentives for compliant agent behavior (Miller 2005, p. 209). Thus, even without comprehensively monitoring agent behavior, the mere ability to observe and evaluate output performance and impose sanctions is enough to discipline an agent. This accountability mechanism can equally be applied to services that are based on ADM systems. Where affected data subjects have the possibility to switch to a different offer, this creates incentives for the operators of an ADM systems to cater to the data subjects' interests.

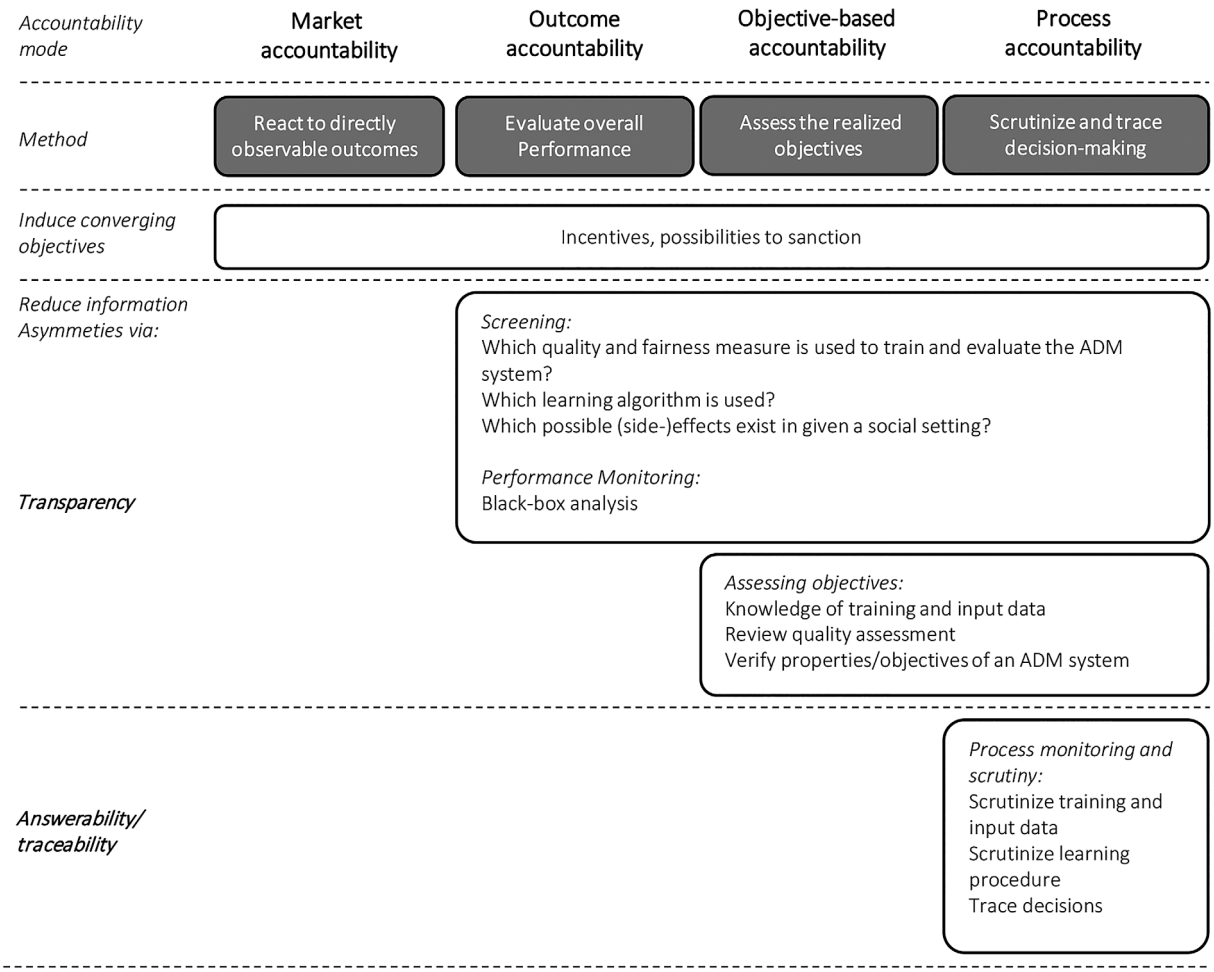

Figure 2 Mechanisms of safeguarding accountability applied to the use of algorithmic decision-making systems.

Secondly, it is possible to go beyond this mode of accountability and to reduce information asymmetries through provisions that allow for screening the agent and more comprehensively monitoring the agent's behavior. With respect to ADM systems, this would mean that rather than trusting that the directly observable decision-making by the ADM system can be evaluated in relation to a principal's interests, one establishes transparency about the overall performance and aggregate impact that is acutely relevant yet not directly observable. This amounts to a kind of *outcome accountability* that involves a more comprehensive performance evaluation than with mere market accountability.

One important instrument for realizing this kind of accountability with regard to ADM systems is the screening of the system based on general information about how the system works and what it does. This may also include obtaining information about the quality measures and parameters that guide the ADM system as well as about the learning algorithm that is used to train the statistical model of the ADM system. Moreover, assessing the overall performance of the ADM system, including not directly observable effects, is made possible through testing in the form black-box analysis (Diakopoulos 2014). This requires that there are interfaces through which the principal can access the system as a black box and see which outputs have been produced based on which inputs. While this method does not actually peak into the black box it can nevertheless reveal patterns from which undesirable behavior of an ADM system can be inferred. This knowledge thus also helps to register unwanted side-effects.

While this way of safeguarding accountability already involves a more systematic and comprehensive assessment of system performance, it still does not go beyond what is commonly called instrumental or outcome

accountability (Patil *et al.* 2014). Like the first mechanism, the focus lies on agent behavior and reliance on sanctioning in the case of undesired performance.

In contrast, the third kind of accountability in Figure 2 aims to achieve transparency with respect to the objectives that an agent realizes. Such an *objective-based accountability* is theoretically possible within the principal-agent paradigm, but it has played a marginal role so far because it is hardly viable with human decision-makers whose authentic values and intentions cannot directly be probed – they ultimately remain a "black box." Trying to get to the intentions of an agent can be seen as a particular form of screening which transcends external aspects of accountability, such as monitoring of behavior, and instead directs attention to internal aspects of accountability (Ebrahim 2003). It aims at ensuring congruent goals and a shared mission of the principal/stakeholder and the agent. Although an assessment of an agent's objectives is not rigorously possible with humans, it is very much a relevant mechanism with regard to ADM systems. Special forms of testing, such as software verification and cryptographic procedures, allow for verifying whether an ADM system has certain properties (Kroll *et al.* 2017).

In a similar vein, assessing the statistical model of an ADM system and inspecting the training data from which it learns altogether realize a sort of transparency that allows for an evaluation of the objectives and the decision quality that a system realizes. With this information, one can calculate and verify not only that a certain quality measure is used, but also whether a certain quality is achieved and how this compares to performance achieved with alternative measures. This objective-based accountability therefore is more stringent than the outcome accountability in making sure that an ADM system realizes certain objectives. More than uncovering hidden action, objective-based accountability aims at hidden characteristics and knowledge regarding an ADM system. It peeks into the black box to make transparent whether the system optimizes the goal that it supposedly does and to check whether the quality measure used really furnishes the best results based on a given standard.

Going even further leads to accountability that thoroughly combines answerability with enforcement: by focusing on the process that leads to decision outcomes. This *process accountability* (Patil *et al.* 2014) also considers whether the agent complies with certain rules of decision-making and it involves the probing or interrogation of the agent. It implies that the agent can be compelled to give an account, to explain and justify her actions for which she can be sanctioned. Process accountability further increases transparency as it makes intelligible or traceable how decisions are and have been produced.

What does this mean specifically for the control of ADM systems? Achieving process accountability with ADM systems means to open up the black box and perform a comprehensive internal review and/or external auditing (see e.g. Bryson & Theodorou 2019; Raji *et al.* 2020). It entails scrutinizing the training and input data (i.e. to trace its generation process and possible error source), the quality assessment (i.e. to probe whether the evaluation is justified), and the learning procedure (i.e. to get insights into the algorithm and the hyperparameters that are used to train the system). Such comprehensive testing could also form the basis for a certification of ADM systems (Matus & Veale 2020).

In its most-demanding form, process accountability allows for tracing and possibly replicating outputs of the ADM system. This may be rather straightforward where certain statistical methods, such as logistic regression, are used for building the ADM model, and where the data is readily available. It may be much more difficult in the case of more complex models such as neural networks, which represent information in ways that are not easily understandable and that may also dynamically change. However, there is ongoing research and development that attempts to make outputs of ADM systems explainable and traceable, even in ways that can be understood by laypersons (Guidotti *et al.* 2018; Sokol & Flach 2020).

Realizing process accountability arguably requires the greatest efforts but it ensures a comprehensive assessment of an ADM system, its design and performance, including the possibility to retrospectively see where things went wrong. Although this accountability mechanism can best minimize agency loss, it may be too much effort for certain uses of ADM systems for which effective provisions that are less demanding might exist.

5. Regulatory requirements for different ADM applications and risks

The various ways of safeguarding accountability described in the previous section are, on the one hand, qualitatively different. On the other hand, they also differ in degree. From purely market-based to process-based

accountability, there is an increase in the strength of accountability constraints as much as in regulatory effort. This means that an efficient regulatory approach toward ADM systems needs to match different instruments for realizing accountability of ADM systems to different degrees of risk. We do so in the following based on a common way of conceptualizing the extent of risk: as an adverse event or effect and the probability of this occurring. These two elements have been used, albeit operationalized differently, to construct risk matrices which allows for matching regulatory provisions to different risks, for example in the financial sector (specifically, the ARROW II model, see Black 2010; MacNeil 2010) and with regard to environmental risks (Black & Baldwin 2012).

As risk matrices are intended to cover a large spectrum of risk settings, they commonly resort to relatively broad categories (such as "medium risk" or "high risk"). However, the purpose of a risk matrix is not to identify concrete and hard thresholds between risk categories, which would also not be useful for regulatory practice (Black & Baldwin 2012, p. 4). The conceptual distinction between risk settings cannot replace a thorough and detailed assessment of concrete cases by regulators. Also, how to deal with a specific source of risk is ultimately a question of social values and risk tolerance that involve a certain malleability and ambiguity – and drawing rigid boundaries between risk categories is therefore hardly appropriate.

Based on these considerations, we distinguish between risk categories primarily to provide a systematic overview that shows how distinctions can be made based on relevant criteria. Our overall goal in doing so is to illustrate how the accountability mechanisms and concrete instruments to establish accountability in the use ADM systems can be matched to different applications depending on the degrees of risk involved.

Conceiving risk in terms of impact and its probability and applying these concepts to algorithmic accountability, the first dimension of impact depends on the potential harm that can follow from ADM operations. It should be noted that the risk of harm is not necessarily *that* an ADM system makes wrong decisions, for example in medical diagnosis and credit default ratings. Wrong decisions will occur, as with human decision-makers, but decisions are and need to be made nonetheless. What is central is to what extent avoidable wrong or bad decisions are minimized – in other words, which objectives are realized and whether certain standards of quality are met. In this sense, the relevant impact is the potential harm from decision consequences in the case of agency loss.

The severity of harm crucially depends on the nature of decision-making – what is decided upon and what are possible decision-outcomes. ADM systems used for consumer recommendations affect individuals' welfare markedly less than ones used for job recruitment or medical interventions. It furthermore matters how many individuals are affected by the decision-making. Even marginal adverse effects due to an ADM system can amount to significant harm caused if a large number of individuals are affected. Also, some ADM systems may produce aggregate, collective adverse effects that cannot easily be reduced to individual impact.

Turning to the likelihood of a negative impact as the second element of risk and transferring it to algorithmic accountability as conceptualized above, this element depends on the potential for agency loss: How much scope is there for an ADM system to realize criteria that diverge from the legitimate expectations and interests of affected data subjects? Thinking in terms of scope instead of probability takes into account that there may be uncertainty and that probabilities are not exactly known – an aspect that has been found wanting in existing risk regulation frameworks (Black 2005, p. 519). Although many uses of ADM systems can easily be assessed in terms of possible decision consequences and potential for undesired performance, a risk evaluation may still involve quite some uncertainty. As this uncertainty forms a possible source of agency loss, it may justify a more comprehensive regulation than in cases where such uncertainty does not exist.

All in all, the potential for agency loss depends on various factors that are linked to information and power asymmetries faced by data subjects. First, it matters whether these have a substantive choice between ADM systems and can easily switch. Where this is the case it reduces the likelihood of agency loss through inducing competitive pressures for service providers to best fulfill the preferences of affected data subjects, whereas the opposite is true in the case of oligopolies. A special case are ADM systems used by the state, from which individuals may not be able to even formally opt out, thus implying a high degree of dependence. Secondly, it matters whether affected data subjects can directly evaluate the ADM performance in relation to their preferences based on perceivable outputs. Thirdly, it makes a difference whether an ADM system is sufficiently complex to make room for unpredictable behavior, that is, through learning in unintended and unforeseen ways. Lastly, complete

automation as part of system design effectively means conferring "authority" to the ADM system, reduces human re-evaluation as a final safeguard, and thus increases the scope for agency loss.

As a general rule, illustrated in Figure 3, the risk for affected data subjects rises as the potential harm from a decision and the scope for the occurrence of agency loss increases (see also Krafft & Zweig 2019). A higher risk, in turn, warrants greater regulatory efforts for ensuring algorithmic accountability. This relationship is of heuristic value, it does not by itself show how different accountability mechanisms and instruments for regulating ADM can be matched to different risk levels. We will elaborate on this link by characterizing different risk settings based on examples that point to relevant criteria and distinctions. It should be noted that the following description illustrates different overall risk levels and corresponding classes of regulation and does not exhaustively deal with all conceivable combinations of potential harm and scope for agency loss. If one of these two risk elements were changed in the examples below while the other remains the same, a lower/higher intensity of regulation becomes more suitable.

5.1. Class 0: Market accountability with or marginal regulation

If the potential harm of decisions is negligible and agency loss is highly unlikely, there is no need for regulation that warrants special transparency and scrutiny of an ADM system. An example for this kind of setting is an ADM system employed to serve customers by automatically matching clothes to customers' individual preferences. A batch of selected clothes is then sent to customers who can keep the clothes they like and send the others back at marginal or no cost. In such a case, harm is negligible; the system may simply not work to realize customers' expectations. The evaluation of the ADM system performance from the point of view of customers is straightforward because it is a matter of taste – they know if the ADM system delivers them outputs that are in their interest, which is very different from situations where it is hard to determine and operationalize the goals that an agent should pursue to maximize one's interests (Osterloh & Frey 2002, p. 114). The customers in the example have a clear standard for evaluating what they get, and ADM performance can directly be evaluated with respect to a data subject's interests. If this furthermore combines with a situation in which data subjects can switch to alternatives and where one can presume competitive incentives among service providers to best cater to these subjects' preferences, regulation is least warranted and market accountability is a suitable solution.

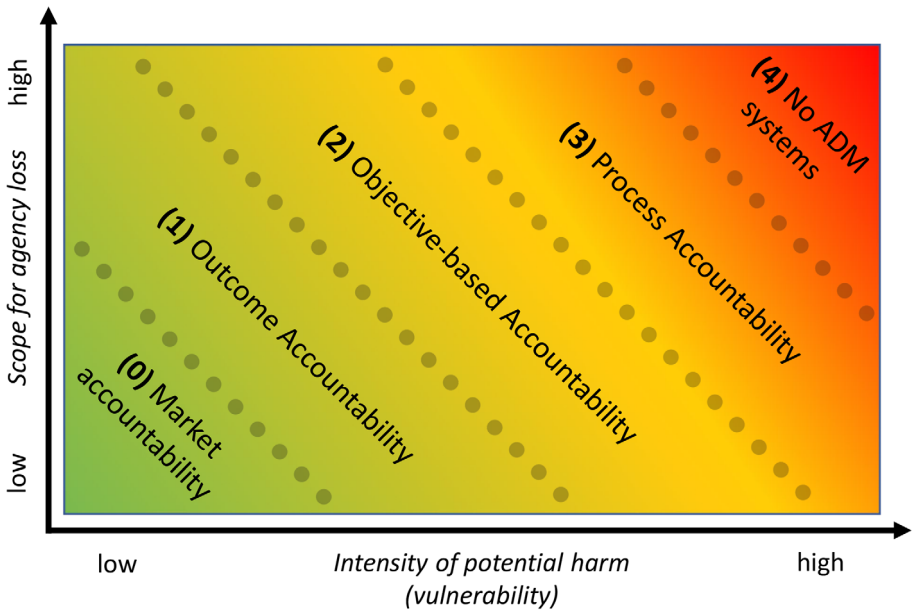

Figure 3 Risk matrix (Adapted from Krafft and Zweig 2019).

5.2. Class 1: Outcome accountability

The situation is different where there is some scope for agency loss and where potential harm is still negligible for individual decisions but can accumulate to non-negligible harm in the aggregate over many decisions (directed at one or many data subjects). An example for this kind of risk is the placement of advertisements as part of a search engine. The provider of a search engine may attempt to bias the placement of ads in favor of its own products or services, which would distort competition and leads to aggregate welfare losses – even though harm of individual decisions (ads displayed to an individual) is negligible. Agency loss is possible in such a case because affected data subjects cannot directly evaluate the performance in relation to their interests. They do not know what an unbiased operating of the ADM system would look like over many decisions. Also, reduced competition means that individuals are comparatively more dependent on the ADM-based service, which decreases incentives of the service provider to cater to the interests of affected data subjects. Given that the potential harm of decisions is not severe, a far-reaching transparency and provisions for looking inside the black box would go too far. A basic screening through disclosure of relevant information and superficial, black-box-testing to monitor behavior of an ADM system, at least by third parties, can already mitigate undesirable aggregate effects that lead to reduced welfare of affected individuals.

5.3. Class 2: Objective-based accountability

If individual decisions can have a significant negative impact on data subjects or a collective of data subjects and there is at notable scope for agency loss it is important to make sure that an ADM system is geared toward realizing objectives that minimize harm. For instance, an ADM system might be used by a company to select applicants for recruiting. This means that the system makes decisions with considerable consequences for people's welfare. Although an applicant has no right to expect to be selected by the ADM system, she does have certain legitimate expectations regarding how the system operates. Specifically, the ADM system in question may discriminate based on sensitive features, such as gender or ethnicity. An equally severe harm can result through ADM systems used in a context where they exert public power, for example through a social network site that uses algorithmic filtering of information streams for millions of people (on this see Just & Latzer 2017). Possible effects ensuing from such a use of ADM systems can even include damage to public goods such as a functioning public sphere and the democratic process. In line with the principle of preventing any undue and uncontrolled use of public power, one can conclude with Tene and Polonetsky (2017, p. 172) that "digital platforms and algorithms should encode the law and widely accepted social values and norms."

In the above-mentioned settings, ADM systems can produce unacceptable, although not existentially severe, harm. This may be the case, for example, because of carelessness and slack resulting from insufficient incentives to avoid harm or because the system learns unexpected, unforeseen decision rules from data while this is not directly apparent for affected data subjects. This kind of constellation warrants regulation that goes beyond the superficial screening and testing. It needs to make sure that certain objectives are transparently realized. In other words, it is then important to be able to make sure whether one is dealing with what Tene and Polonetsky (2017, p. 126) have called "policy-directed algorithms," which are engineered for the purpose of avoiding or correcting biases and advancing specific objectives.

With human decision-makers, more far-reaching, procedural scrutiny may well be appropriate as their motives will ultimately remain black-boxed. With ADM systems, however, it is possible to examine whether an ADM system has certain properties and conforms to certain criteria without already going one step further and introducing procedural accountability requirements. If an ADM system has been adequately designed and prepared beforehand, it is possible to perform tests which can ensure that a certain fairness requirement is met (Kroll *et al.* 2017). It is thus possible to make the realized objectives and decision quality transparent without completely illuminating the "black box."

5.4. Class 3: Process accountability

Where potential harm of individual decisions is severe and there is notable scope for the ADM system to operate in a way that diverges from the legitimate expectations and interests of affected data subjects without them knowing, there are grounds for very strict regulation that aims at mitigating risk as much as possible – that is,

detecting and avoiding any way in which avoidable harmful decision could be made. This constellation is similar to validation of aircraft autopilots for which rigorous design and testing procedures have been developed which also allows for tracing decisions of the system itself. A different example would be the application of ADM systems for medical diagnostics, such as the detection of certain forms of breast cancer from tomographic imaging. Both a wrong negative (no treatment despite lethal illness) and a wrong positive (possibly mastectomy although no illness present) have far-reaching consequences. Even when presuming that the interest of those designing and offering an ADM application for that purpose is aligned with those about whom decisions are made, the ADM system might still have an undesirable bias or learn decision rules in an unexpected fashion – thus making avoidable wrong decisions with potentially fatal consequences.

Indeed, it is known that ADM systems used for pattern recognition in tomography have a better recall (classifying as positive those cases which, in fact, are positive) than humans, but they produce markedly more false positives – which is why human experts are required as a corrective to at least check the positive decisions, which are a much more manageable number than all cases (Fry 2018). To make sure that risk is mitigated as much as possible in such an application of an ADM system, comprehensive audits are needed that examine the construction of the ADM system, that entail a thorough testing and validation, and that can make decisions traceable. This also allows for seeing retrospectively what may have gone wrong. In order to be able to reach an acceptable level of transparency and traceability, it may furthermore be required to only use certain statistical models that are still interpretable for humans (Rudin 2019).

It should be noted that while such a comprehensive auditing aims to make sure that an ADM system conforms to certain criteria and quality standards, it does not as such resolve the question what suitable criteria are in the first place. This question can become especially salient and thorny where difficult trade-offs between social values are involved. Such trade-offs cannot easily be decided by regulators and some scholars have noted that there may be strong reasons for involving stakeholders to resolve these issues (see e.g. Lepri *et al.* 2018; Veale *et al.* 2018). For instance, with regard to recidivism risk assessments in criminal justice, it is not per se clear how much more severe a decision outcome is compared to another: wrongly predicting someone as recidivist and detaining them (false positive) or wrongly predicting someone as a low risk and letting them free (false negative). This question amounts to choosing a suitable quality measure that is to be optimized.[4]

5.5. Class 4: No use of ADM systems

Where fundamental values are affected and there are unacceptable value trade-offs, the use of an ADM system may quickly reach a point at which a society and regulators see it as inadmissible – even if safeguards can be implemented. The potential harm of decisions might be so severe and/or affect so many people while there is considerable scope for an unintended and undesirable operation of an ADM system that it is intolerable. If, for instance, the complexity of such an ADM system makes it barely predictable while an efficient application of that system demands automation, it will hardly be acceptable. Where such a line will be drawn is subject to debate. For example, there has recently been controversy over ADM-based political microtargeting via online advertisements – regarding how to regulate it and whether to ban it altogether. Another highly controversial application of ADM is lethal autonomous weapons that may make life-and-death decisions without oversight. In any case, as with certain technologies that carry significant risks, there may be reasons not to use ADM systems at all.

6. Conclusion

The present paper contributes to the existing literature with a framework that differentiates regulatory requirements for variable implementations of ADM systems. We have drawn on the principal-agent model as a conceptual lens to systematically examine the accountability issues arising with the adoption of ADM systems. The studied accountability issues are those from the perspective of data subjects who provide input – often including their personal data – to an ADM system that produces outputs in the form of decisions made for and/or about these data subjects.

A major challenge from this perspective are information asymmetries. These create possibilities for agency loss in the sense that an ADM system operates in a way that goes against the legitimate expectations and interests

of affected data subjects. The principal-agent literature distinguishes several dimensions of that asymmetry which can be applied to the relation between data subjects and ADM systems. These may, first of all, exhibit hidden qualities and objectives as they may perform in undesirable ways and realize criteria that go against the legitimate interests of data subjects without them being aware of this. Such objectives may be designed into the system, but they can also be the unintended result of an ADM system learning from data. Secondly, data subjects may face the problem of hidden knowledge/information as they do not know which information and analytical insights the system uses. Lastly, hidden actions can occur with ADM systems due to their usually unobtrusive operations as well as their opacity and complexity.

Based on the principal-agent model, we have furthermore distinguished general ways of dealing with the accountability problems resulting from those asymmetries and have applied them to ADM systems. A first variant is to only rely on pure market accountability, that is, substantive choices between offers inducing incentives for hosts and designers of ADM systems to make them work in the interest of affected data subjects. Secondly, one may impose relatively superficial information disclosure requirements and transparency provisions, which may also involve black-box testing as a way to monitor and assess aggregate performance of an ADM system. The third mechanism described above is a theoretical possibility in the principal-agent framework but is hardly viable with humans and only becomes relevant with machines realizing decision-making and services. Specifically, more far-reaching transparency provisions through special forms of testing aim at ascertaining which objectives are realized by an ADM system – they allow for peaking inside the "black box." Finally, comprehensive testing and auditing can be used to inspect the construction of the system and its data processing, and to make decisions traceable.

These are not only different kinds of mechanisms. They also differ in degree regarding the intensity of regulatory intervention. Similar to contributions on risk regulation, we have matched regulatory provisions to different applications of ADM systems based on the risks involved. This risk depends on the potential harm of decisions and the scope for agency loss, which in turn results from the specific, socially embedded implementation of an ADM system. Although there are no clear-cut thresholds between risk categories, it is nonetheless important not to lose sight of the big picture since applications of ADM differ widely in their impact on society and the risks they entail. The framework formulated above contributes to a broader and systematic view on ADM systems. It offers an overview of how different means for safeguarding algorithmic accountability which have been proposed in the literature can be systematized and linked to a broad range of ADM system applications based on relevant criteria; and it does so with a firm conceptual footing that has already been applied to accountability problems in other areas.

Indeed, varying requirements of scrutiny and transparency already exist with regard to human decision-making. This will similarly have to be realized with ADM systems as these increasingly mediate social relations. The core accountability challenges are not new, but they manifest themselves in new ways and call for new instruments, including special technological solutions. The formulated framework also underscores that, as with human decision-making and accountability, ethics is not enough. While there is growing literature on algorithm ethics and one can discern a tendency by private as much as state organizations to adopt ethical norms concerning the use of ADM systems, the framework above suggests that there have to be effective regulatory instruments in place that allow for transparency and answerability where ADM systems are adopted. At the same time, these need to be chosen carefully in order to avoid costs and overregulation that may stifle beneficial uses of ADM systems.

Acknowledgments

We thank the anonymous reviewers and the editors of *Regulation & Governance* for their helpful comments and suggestions. The paper also owes to Georg Wenzelburger and Philipp Bird.

This research has been conducted within the project "Deciding about, by, and together with algorithmic decision making systems," funded by the Volkswagen Foundation. Open access funding enabled and organized by Projekt DEAL.

Endnotes

[1] Strictly speaking, it is operationalized behavior, which can be processed by a machine.

[2] An ADM system may well be unacceptable on the grounds of a right to privacy or other personal rights because it depends on far-reaching surveillance structures. However, this touches on regulatory issues different from the ones we are concerned with given our focus on decision-making.

[3] Some of these applications also time involve explicit decision-making about the data subjects, as part of the service. This is the case, for example, with medical diagnosis tools, with which users can find out something about themselves. Other applications that, for example, involve a profiling of individuals instead make implicit decisions about them.

[4] A similar problem emerges with choosing a fairness measure (Berk *et al.* 2018).

References

Abbott KW, Levi-faur D, Snidal D (2017) Theorizing Regulatory Intermediaries: The RIT Model. *The Annals of the American Academy of Political and Social Science* 670, 14–35. https://doi.org/10.1177/0002716216688272.

Ananny M, Crawford K (2018) Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. *New Media & Society* 20, 973–989. https://doi.org/10.1177/1461444816676645.

Barocas S, Selbst AD (2016) Big Data's Disparate Impact. *California Law Review* 104, 671–732.

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 (online first)*, 1–42. https://doi.org/10.1177/0049124118782533.

Binns R (2017) Algorithmic Accountability and Public Reason. *Philosophy & Technology* 31, 543–556. https://doi.org/10.1007/s13347-017-0263-5.

Black J (2005) The Emergence of Risk-Based Regulation and the New Public Risk Management in the United Kingdom. *Public Law* 4: 512–548.

Black J (2010) Risk-Based Regulation: Choices, Practices and Lessons Being Learnt. In: OECD (ed) *Risk and Regulatory Policy: Improving the Governance of Risk*. OECD Reviews of Regulatory Reform, pp. 185–236. OECD, Paris. https://doi.org/10.1787/9789264082939-en.

Black J, Baldwin R (2012) When Risk-Based Regulation Aims Low: Approaches and Challenges: Aiming Low: Approaches and Challenges. *Regulation & Governance* 6, 2–22. https://doi.org/10.1111/j.1748-5991.2011.01124.x.

Brauneis R, Goodman EP (2018) Algorithmic Transparency for the Smart City. *The Yale Journal of Law and Technology* 20, 103–175. https://doi.org/10.2139/ssrn.3012499.

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. https://doi.org/10.1093/ijlit/eay017.

Bryson JJ, Theodorou A (2019) How Society Can Maintain Human-Centric Artificial Intelligence. In: Toivonen M, Saari E (eds) *Human-Centered Digitalization and Services*, Vol. 19, pp. 305–323. Springer Singapore, Singapore. https://doi.org/10.1007/978-981-13-7725-9_16.

Bygrave LA (2019) Minding the machine v2.0: The EU General Data Protection Regulation and Automated Decision Making. In: Algorithmic Regulation (ed) *Karen Yeung and Martin Lodge*, pp. 248–262. Oxford University Press, Oxford.

Diakopoulos N (2014) *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*. Columbia University. https://doi.org/10.7916/d8zk5tw2.

van Drunen MZ, Helberger N, Bastian M (2019) Know your Algorithm: What Media Organizations Need to Explain to Their Users about News Personalization. *International Data Privacy Law* 9, 1–16. https://doi.org/10.1093/idpl/ipz011.

Ebrahim A (2003) Making Sense of Accountability: Conceptual Perspectives for Northern and Southern Nonprofits. *Nonprofit Management and Leadership* 14, 191–212. https://doi.org/10.1002/nml.29.

Edwards L, Veale M (2017) Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking for. *Duke Law & Technology Review* 16, 18–84.

Eyert F, Irgmaier F, Ulbricht L (2022) Extending the framework of algorithmic regulation. The Uber case. *Regulation & Governance* 16(1), 23–44.

Fry H (2018) *Hello World: How to Be Human in the Age of the Machine*. Doubleday, London.

Gellert R (2022) Comparing Definitions of Data and Information in Data Protection Law and Machine Learning: A Useful Way Forward to Meaningfully Regulate Algorithms? *Regulation & Governance* 16(1), 156–176.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* 51, 1–42. https://doi.org/10.1145/3236009.

Heremans T (2012) *Professional Services in the EU Internal Market: Quality Regulation and Self-Regulation*. Hart, Oxford.

Hildebrandt M (2008) Defining Profiling: A New Type of Knowledge? In: Hildebrandt M, Gutwirth S (eds) *Profiling the European Citizen*, pp. 17–45. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-1-4020-6914-7_2.

Hildebrandt M (2016) Law as Information in the Era of Data-Driven Agency: Law as Information. *Modern Law Review* 79, 1–30. https://doi.org/10.1111/1468-2230.12165.

Hill CWL, Jones TM (1992) Stakeholder-Agency Theory. *Journal of Management Studies* 29, 131–154. https://doi.org/10.1111/j.1467-6486.1992.tb00657.x.

Hirschmann AO (1970) *Exit, Voice and Loyalty: Responses to Decline in Firms, Organizations and States*. Harvard University Press, Cambridge.

Hölmstrom B (1979) Moral Hazard and Observability. *The Bell Journal of Economics* 10, 74–91.

Hölmstrom B, Milgrom P (1987) Aggregation and Linearity in the Provision of Intertemporal Incentives. *Econometrica* 55, 303–328. https://doi.org/10.2307/1913238.

Jensen MC, Meckling WH (1976) Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics* 3, 305–360. https://doi.org/10.1016/0304-405X(76)90026-X.
Just N, Latzer M (2017) Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet. *Media, Culture & Society* 39, 238–258. https://doi.org/10.1177/0163443716643157.
Kerwer D (2005) Holding Global Regulators Accountable: The Case of Credit Rating Agencies. *Governance* 18, 453–475. https://doi.org/10.1111/j.1468-0491.2005.00284.x.
Koene A, Clifton CW, Hatada Y *et al.* (2019) *A Governance Framework for Algorithmic Accountability and Transparency*. European Parliamentary Research Service, Brussels.
Koops B-J (2013) On Decision Transparency, or How to Enhance Data Protection after the Computational Turn. In: Hildebrandt M, de Vries K (eds) *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, pp. 196–220. Milton Park, Routledge.
Krafft TD, Zweig K (2019) Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse. *Ein Regulierungsvorschlag*. [Last accessed 14 Oct 2020.] Available from URL: https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf
Kroll JA, Huey J, Barocas S *et al.* (2017) Accountable Algorithms. *University of Pennsylvania Law Review* 165, 633–705.
Lane J-E (2007) *Comparative Politics: The Principal-Agent Perspective*. Routledge, Milton Park. https://doi.org/10.4324/9780203935545.
Lanzing M (2018) "Strongly Recommended" Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies. *Philosophy & Technology* 32, 549–568. https://doi.org/10.1007/s13347-018-0316-4.
Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy & Technology* 31, 611–627. https://doi.org/10.1007/s13347-017-0279-x.
Lodge M (2004) Accountability and Transparency in Regulation: Critiques, Doctrines and Instruments. In: Jordana J, Levi-Faur D (eds) *The Politics of Regulation*, pp. 124–144. Cheltenham: Edward Elgar Publishing.
Lupia A (2003) Delegation and Its Perils. In: Strøm K, Bergman T, Müller WC (eds) *Delegation and Accountability in Parliamentary Democracies*, pp. 33–54. Oxford University Press, Oxford; New York.
MacNeil I (2010) Risk Control Strategies: An Assessment in the Context of the Credit Crisis. In: MacNeil I, O'Brien J (eds) *The Future of Financial Regulation*, pp. 141–160. Hart, Oxford; Portland, OR.
Malgieri G (2019) Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards" in the National Legislations. *Computer Law & Security Review* 35, 1–26. https://doi.org/10.1016/j.clsr.2019.05.002.
McCubbins MD, Schwartz T (1984) Congressional Oversight Overlooked: Police Patrols Versus Fire Alarms. *American Journal of Political Science* 28, 165–179.
Miller GJ (2005) The Political Evolution of Principal Agent Models. *Annual Review of Political Science* 8, 203–225. https://doi.org/10.1146/annurev.polisci.8.082103.104840.
Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* 3, 1–21. https://doi.org/10.1177/2053951716679679.
Nalebuff BJ, Stiglitz JE (1983) Prizes and Incentives: Towards a General Theory of Compensation and Competition. *The Bell Journal of Economics* 14, 21. https://doi.org/10.2307/3003535.
Nemitz P (2018) Constitutional Democracy and Technology in the Age of Artificial Intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 1–14. https://doi.org/10.1098/rsta.2018.0089.
Osterloh M, Frey BS (2002) Does Pay for Performance Really Motivate Employees. In: Neely AD (ed) *Business Performance Measurement: Theory and Practice*, pp. 107–122. Cambridge University Press, Cambridge.
Pasquale F (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press, Cambridge, MA; London.
Patil SV, Vieider F, Tetlock PE (2014) Process Versus Outcome Accountability. In: Patil SV, Vieider F, Tetlock PE (eds) *The Oxford Handbook of Public Accountability*, pp. 69–89. Oxford University Press, Oxford.
Pratt JW, Zeckhauser R (1991) Principals and Agents: An Overview. In: Pratt JW, Zeckhauser R (eds) *Principals and Agents: The Structure of Business*. Research Colloquium, pp. 1–36. Harvard Business School Press, Boston, MA.
Raji ID, Smart A, White RN *et al.* (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44. ACM, Barcelona. https://doi.org/10.1145/3351095.3372873.
Robinson H, Carrillo P, Anumba CJ, Patel M (2010) *Governance and Knowledge Management for Public-Private Partnerships*. Wiley-Blackwell, Chichester; Malden, MA.
Rudin C (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x.
Saam NJ (2007) Asymmetry in Information Versus Asymmetry in Power: Implicit Assumptions of Agency Theory? *The Journal of Socio-Economics* 36, 825–840. https://doi.org/10.1016/j.socec.2007.01.018.
Sappington DEM (1991) Incentives in Principal-Agent Relationships. *Journal of Economic Perspectives* 5, 45–66. https://doi.org/10.1257/jep.5.2.45.
Saurwein F, Just N, Latzer M (2015) Governance of Algorithms: Options and Limitations. *Info* 17, 35–49. https://doi.org/10.1108/info-05-2015-0025.
Schiffman LG, Kanuk LL, Hansen H (2012) *Consumer Behaviour: A European Outlook*, 2nd edn. Pearson Financial Times/Prentice Hall, Harlow; New York.
Sherman R (2011) *The Regulation of Monopoly*. Cambridge University Press, Cambridge.

Sokol K, Flach P (2020) Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 56–67. ACM, Barcelona. https://doi.org/10.1145/3351095.3372870.

Tene O, Polonetsky J (2017) Taming the Golem: Challenges of Ethical Algorithmic Decision-Making. *North Carolina Journal of Law & Technology* 19, 125–173.

Ulbricht L, Yeung K (2022) Algorithmic Regulation: A Maturing Concept for Investigating Regulation of and Through Algorithms. *Regulation & Governance* 16(1), 3–22.

Veale M (2020) The Provenance of Trained Machine Learning Models: Will Tomorrow's AI Systems Need Fairtrade Certification? *Regulation & Governance*, forthcoming.

Veale M, Van Kleek M, Binns R (2018) Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems – CHI '18*, pp. 1–14. ACM Press, Montreal. https://doi.org/10.1145/3173574.3174014.

Warren ME (2014) Accountability and Democracy. In: Bovens M, Goodin RE, Schillemans T (eds) *The Oxford Handbook of Public Accountability*, pp. 39–54. Oxford University Press, Oxford.

Watt J, Borhani R, Katsaggelos AK (2020) *Machine Learning Refined: Foundations, Algorithms, and Applications*, Second edn. Cambridge University Press, New York.

Weingast BR, Moran MJ (1983) Bureaucratic Discretion or Congressional Control? Regulatory Policymaking by the Federal Trade Commission. *Journal of Political Economy* 91, 765–800. https://doi.org/10.1086/261181.

Weiss MD (1995) Information Issues for Principals and Agents in the "Market" for Food Safety and Nutrition. In: Caswell JA (ed) *Valuing Food Safety and Nutrition*, pp. 69–82. Routledge, Milton Park.

Wieringa M (2020) What to Account for when Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 1–18. ACM, Barcelona. https://doi.org/10.1145/3351095.3372833.

Yeung K (2017a) "Hypernudge": Big Data as a Mode of Regulation by Design. *Information, Communication & Society* 20, 118–136. https://doi.org/10.1080/1369118X.2016.1186713.

Yeung K (2017b) Algorithmic Regulation: A Critical Interrogation: Algorithmic Regulation. *Regulation & Governance* 12, 505–523. https://doi.org/10.1111/rego.12158.

Zweig KA, Fischer S, Lischka K (2018) *Wo Maschinen irren können. Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung*. Bertelsmann Stiftung, Gütersloh.

