# Beyond explainability: justifiability and contestability of Algorithmic Decision Systems

Clément Henin, Daniel Le Métayer

# Beyond explainability: justifiability and contestability of Algorithmic Decision Systems

Clément Henin[1,2] and Daniel Le Métayer[1]

[1] Univ Lyon, Inria, INSA Lyon, CITI, Villeurbanne, France {clement.henin, daniel.le-metayer}@inria.fr
[2] École des Ponts ParisTech, Champs-sur-Marne, France

**Abstract.** In this paper, we point out that explainability is useful but not sufficient to ensure the legitimacy of algorithmic decision systems. We argue that the key requirements for high stakes decision systems should be justifiability and contestability. We highlight the conceptual differences between explanations and justifications and suggest different ways to operationalize justifiability and contestability.

**Keywords:** challenge · contestation · justification · explanation · machine learning · evidence.

## 1 Introduction

Algorithms are increasingly used to support decision making. The nature of these Algorithmic Decision Systems (hereinafter "ADS") varies: some of them rely on machine learning while others do not; some of them involve a form of interaction with human users while others are entirely automatic; some of them are intended for professionals while others are aimed at the general public. Regardless of these differences, ADS are often involved in decisions that can have a significant impact on people: access to credit, employment, medical treatment, judicial sentences, etc. Entrusting ADS to make or to influence such decisions raises a variety of legal, ethical, political and technical issues. If these issues are not properly addressed, the expected benefits of these systems may be offset by unacceptable risks for individuals (discrimination, loss of autonomy, etc.), the economy (unfair practices, limited access to markets, etc.) and society as a whole (manipulation, threat to democracy, etc.). Broad requirements such as transparency, fairness and accountability are often presented as ways to limit these risks but they are generally ill defined, seldom required by law and difficult to implement[19, 42]. For example, the terms "transparency", "explanation" and "justification" are frequently used without precise definition, sometimes interchangeably, sometimes with different meanings. On the technical side, a lot of effort has been put on bias reduction, accuracy improvement and the generation of different types of explanations with or without access to the code of the system ("white box" versus "black box" methods) [2, 5, 36, 39]. However, the integration of these technical results within a responsible approach for the development and deployment of ADS is rarely discussed.

The aim of this paper is to contribute clarifying the debate about ADS and suggesting conceptual and technical instruments to foster a development process addressing the potential risks that ADS may pose. First, we propose precise definitions for the main concepts used in this paper and emphasize the distinction between explanations and justifications (Section 2). Then, we argue that the essential conditions to ensure the legitimacy of high stakes ADS are justifiability and contestability rather than explainability (Section 3). Justifiability and contestability have not received as much attention as explainability in the computer science community so far. Nevertheless, we show that they can be put into practice at different stages of the life-cycle of an ADS and suggest ways to operationalize them in Section 4. We present related work in Section 5 and conclude with a more general discussion in Section 6.

## 2 Terminology

In this section, we define the terminology used in this paper. This terminology has not yet stabilized in the literature and the same words are sometimes used with different meanings by different authors. We discuss these variations in Section 5.

**Algorithmic decision system (ADS):** We use the expression algorithmic decision system (ADS) rather than algorithm to stress the fact that algorithms "should be studied in a general setting that includes their parameters, context of use and, if they rely on machine learning, their training data" [19]. We do not make any assumption on the techniques used to implement ADS and the type of interactions they have with their environment. The impact of these aspects on the requirements put on ADS is analysed in Section 3.6.

**Explanation, explainable, explainability:** the goal of an explanation is to make it possible for a human being (designer, user, affected person, etc.) to *understand* (a specific outcome or the whole system). For example, an explanation for a bank loan application rejection could be that the number of outstanding loans of the applicant is too high (e.g. greater than a given threshold). This information helps to understand the logic of the ADS.

**Justification, justifiable, justifiability**: the goal of a justification is to convince that a decision is *good* (or adequate, appropriate)[3]. A justification of the above loan rejection could be that applications with many outstanding loans have a high probability to lead to credit defaults, which is a risk that the bank wants to reduce. Another justification could be that banking law prohibits granting new loans when the number of outstanding loans of the applicant exceeds a given threshold.

**Contestation, contestable, contestability**: the goal of a contestation is to convince that the decision is *bad* (or inadequate, inappropriate). A contestation of the above decision could rely on the fact that the applicant has a significant amount of saving.

[3] Or, more generally, that the outcomes of an ADS are appropriate (global justification).

**Account, accountable, accountability**: according to Reuben Binns [12], "a party A is accountable to a party B with respect to its conduct C, if A has an obligation to provide B with some justification for C, and may face some form of sanction if B finds A's justification to be inadequate." The accounts are the pieces of information that must be provided by the accountable party. We use accountability in the same sense as Binns and therefore consider that justifications are an essential part of the accounts. Since justifications may not be considered "adequate" (using Binns' terminology), contestability is also a key requirement for accountability. This point is further discussed below.

**Legitimate, legitimacy**: Many definitions of legitimacy have been proposed by political scientists, lawyers and philosophers. We use Mark Suchman's definition here [68], which is general enough to apply to ADS [72]. Suchman defines legitimacy as "a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions."

Our definitions make a clear distinction between explanations and justifications - two terms that are used interchangeably in some papers. Explanations are transfers of knowledge (from the ADS to the explainee), they are descriptive and intrinsic in the sense that they only depend on the system itself[4]. In contrast, justifications are normative and extrinsinc in the sense that they depend on a reference according to which the adequacy or appropriateness of the outcomes can be assessed. Indeed, in order to claim that an outcome is good (or adequate, appropriate) it is necessary to refer to an independent definition of what a good outcome is. In this paper, we use the word "norm" to denote this external reference, with no legal connotation: a norm can be a legal requirement but it can also be a corporate objective or an ethical principle for example. In the above example, the first justification is based on a corporate objective (minimization of the risks for the bank) while the second one relies on a legal norm (banking law).

Even if they often support each other, explanations and justifications have different goals and should not be conflated: a user can understand the logic leading to a particular outcome without agreeing on the fact that this outcome is good; vice versa, he/she may want to contest an outcome (being convinced that it is bad) without knowing or understanding the logic behind the algorithm. The distinction between the two notions is best illustrated by Mireille Hildebrandt [40]: "If the decisions of an automated machine learning application indirectly discriminate on the basis of gender or race they may qualify as prohibited discrimination; explaining why the system so decided may be interesting but will

[4] This is also the case for "causal explanations": even though the notion of cause is very complex and it is used with a variety of different meanings in the literature, causal explanations are generally based on relations between ADS inputs and outputs, without reference to any external norm [3].

not legally justify the decision. A decision of an automated system should be justifiable independently of how the system came to its conclusion."[5]

A contestation can be seen as the dual of a justification: its goal is to convince that the decision is bad while the goal of a justification is to convince that the decision is good. They are also dual in the sense that the source (or issuer) of a justification is the operator (or designer) of the ADS whereas the source of a contestation is a user of the ADS or a person affected by its decisions. Justifications alone would lead to a unilateral process whereas contestations introduce a dialectic, conversational process.

Last but not least, all the notions introduced here can take two forms: local or global, meaning that they can apply to individual decisions or to the overall ADS. For example, different types of explanations can be provided about a given outcome[6] or about the global logic of the ADS[7]. Similarly, legitimacy can be considered at the global level (for example in the context of an *ex-ante* algorithmic impact assessment, as discussed in Section 6), or regarding a specific decision.

## 3 Why contestability should be a requirement

A lot of progress has been made in artificial intelligence (AI) during the last decades and the promoters of these technologies, in particular machine learning (ML), have raised high expectations about their applications in many areas. However, academics, NGOs and civil society in general have expressed concerns and fears about the impacts of the pervasive deployment of AI, in particular in the context of ADS. As discussed in [20], this debate is often clouded by the mix of arguments of a totally different nature such as issues about the legitimacy of the very purpose of the system and more specific questions about technical choices made for its implementation. In this section, we start with general concerns raised about the deployment of ADS and use them as the driving motivation to define the requirements that these systems should meet.

### 3.1 Justifiability and contestability are necessary conditions for the legitimacy of algorithmic decisions

The fact that ADS are increasingly involved in many everyday decisions (shopping recommendations, access to information, targeted advertisements, etc.) but also in more important instances (health-care coordination, diagnosis, lawsuits, credit applications, job applications, etc.) has led some authors to raise the

[5] Mireille Hildebrandt takes as an illustration the example of courts of justice: "When a court decides a case, it cannot justify its decision by spelling out the heuristics of the judge(s) involved, such as their political preferences, what they had for breakfast or how they prepared the case."

[6] For example, the factors (input values) that had the strongest impact on the outcome.

[7] For example, in the form of a decision tree or a list of rules.

risks of "algocracy" [27] or "algorithmic governmentality" [66]. As stated by Antoinette Rouvroy [66], "this 'algorithmic governmentality', and its self-enforcing, implicit, statistically established norms emanating, in real time, from digitalized reality, contrasts with 'political governmentality', and the imperfectly enforced, explicit, deliberated, character of laws resulting from time consuming political deliberation." Ari Ezra Waldman [72] concurs, stating that "The result is a technologically driven decision-making process that seems to defy interrogation, analysis, and accountability and, therefore, undermines due process. This should make algorithmic decision-making an illegitimate source of authority in a liberal democracy."

The key question regarding the deployment of ADS is therefore the requirements that should be imposed to ensure their legitimacy or the legitimacy of the decisions based on their outcomes[8]. Kees Van Kersbergen and Frans Van Waarden refer to accountability as a condition for input legitimacy[9] [45] and Steven Bernstein highlights the importance of justifications and contestations to ensure legitimacy [9].

### 3.2 Justifications and contestations must rely on norms

The need to provide ways to challenge an ADS is recognized by law but legal requirements are generally expressed as ill-defined rights to obtain explanations or human intervention. For example, according to Recital 71 of the GDPR, when an automated processing is used to make a decision about a person[10], this person has "the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision." [11] As regards ethics, the Ethics Guidelines for Trustworthy AI published by the High Level Expert Group on Artificial Intelligence (HLEG-AI) [42] also put forward explainability as one of their four main principles[12]. This emphasis is motivated as follows: "Explicability is crucial for building and maintaining users' trust in AI systems. [...] Without such information, a decision cannot be duly contested." It is interesting to notice that both in the GDPR and in the HLEG-AI guidelines, explainability requirements are

[8] Global versus local legitimacy.

[9] A distinction is usually drawn between two ways of ensuring legitimacy, called respectively input legitimacy and output legitimacy. Input legitimacy focuses on procedural aspects (policies and participation of the stakeholders) while output legitimacy concerns the effectiveness of the policies (quality of the results) [33]. John Danaher uses the expressions "instrumentalism" and "proceduralism" to denote respectively output and input legitimacy [27].

[10] More precisely "a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her."

[11] It should be noted however that the interpretation of the GDPR regarding explainability requirements has stimulated some debate among law experts [54, 71].

[12] Together with respect for human autonomy, prevention of harm and fairness.

immediately followed (and implicitly motivated) by the need to allow affected people to contest decisions.

However, as argued in the previous section, explanations are not sufficient to justify or to contest decisions. This point is forcefully argued by Mireille Hildebrandt [40]: "we must not allow the discourse of explainability to stand in the way of the question whether a decision is legally justified, which requires a specific type of legal reasons. Explanation in itself does not imply justification, and justification does not always require an explanation of the underlying logic of the decision system."

In contrast with explanations, justifications rely on norms, that is to say requirements that are generally outside the algorithmic system[13]. Law is obviously a prime example of norm, but, as discussed below, different sources of norms can apply to a given system. The key point is that justifications make it possible to avoid the "self-production" of norms (norms emanating from the system itself, without external reference or control) pointed out by Antoinette Rouvroy [66, 67].

Norms are varied, they can have different sources of legitimacy and can be expressed in different ways (e.g. through law or jurisprudence for legal norms). When several norms apply, they may be in tension, or even in contradiction. In some cases, it is possible to rely on priority rules to establish precedence of a norm over another one (e.g., international law usually prevails over domestic law, constitution prevails over ordinary laws, which prevail over decrees, etc.); in other cases, such rules may not exist and the conflicts between them must be solved on a case by case basis.

### 3.3 Eliciting applicable norms is, in itself, a sound discipline

The elicitation of norms justifying the outcomes of an ADS and their use in a particular context is not only useful to enable contestability: it is also a sound discipline to increase the chances that the ADS has a legitimate objective and is used in an appropriate way. In practice, critical assumptions about the ultimate goal of an ADS and the reasons to believe that it is well-suited to address them, are often left implicit or unclear. This seems to be a typical cause of misuse of these systems and mistrust about them. An instructive example is the use of ADS by certain courts of justice to assess the risk of recidivism of defendants or detainees (or failure to appear in the case of pretrial assessments). As stressed by Angèle Christin and her co-authors [23], the "more problematic is the theory of justice implicitly embedded in the algorithms. Punishment is usually said to have four main justifications: retribution[14], deterrence[15], incapacitation[16],

[13] We discuss in Section 5 contexts, such as autonomous agents, in which an ADS can incorporate certain norms.

[14] The punishment must fit the crime and be proportionate to the severity of the infraction.

[15] The punishment discourages people from committing crimes.

[16] The punishment positively prevents someone from offending, for example through imprisonment.

and rehabilitation[17]. Risk-assessment tools emphasize one major justification at the detriment of the others: incapacitation." The authors conclude with key questions about the role that algorithms should play in criminal justice decision-making, how to ensure that they are used appropriately and how to challenge their decisions. [18]

As the example of fairness teaches us, the choices of values embedded in ADS should be subject to a broad debate, which can be facilitated by the elicitation of the (possibly conflicting) norms at stake. Indeed, fairness can be defined in many ways and some of these definitions are incompatible with one another and may be in tension with the accuracy objective. As noticed by Richard Berk and his co-authors [8], the trade-offs may be challenging and application dependent. For example, technically speaking, age and gender are critical factors to predict recidivism (young men having the highest level of risk) as well as to anticipate the development of certain diseases. However, the use of this information by a judge could be considered as a form of discrimination while it goes without saying that a doctor would take it into account to make a decision about a patient. As Richard Berk and his co-authors conclude, "in the end, it will fall to stakeholders – not criminologists, not statisticians and not computer scientists – to determine the trade-offs. How many unanticipated crimes are worth some specified improvement in conditional use accuracy equality ? [...] These are matters of values and law, and ultimately, the political process. They are not matters of science." In a sense, this shift from facts to values is precisely the shift from explanations, which reflect the logic of the ADS, to justifications, which refer to external, debatable norms.

### 3.4 Justifications and contestations are essential parts of accountability

As suggested in Section 3.1 and Section 3.2, accountability is a key component of legitimacy. The significance of accountability has been highlighted by many authors. For example, according to Black [17], accountability relationships "are a critical element in the construction and contestation of legitimacy claims by both regulators and legitimacy communities, as they are the means by which legitimacy communities seek to ensure that their legitimacy claims are met, and that their evaluations of the legitimacy of regulators are valid." In the same vein, Mark Bovens [18] sees accountability as "a route through which pragmatic and moral/normative legitimacy claims in particular are validated" and posits that "public accountability is indirectly of importance because ultimately, it can help

[17] Which emphasizes instead the potential recovery of offenders and their inclusion in the social body.

[18] John Monahan and Jennifer L. Skeem argue in the same direction in their analysis of risk assessment in criminal sentencing [59]. Chelsea Barabas and her co-authors go further, suggesting that machine learning should not be used for risk prediction but for risk mitigation because empirical analysis has demonstrated that it is "ineffective at lowering near-term risks (failure to appear and new criminal activity) and long-term recidivism rates."

to ensure that the legitimacy of the public administration remains intact or is increased."

The definition provided in Section 2 (taken from [12]) characterizes accountability as the obligation to provide justifications and the possibility to challenge them[19]. In the same spirit, Mark Bovens [18] defines accountability as "the obligation to explain and justify conduct" and considers that it "usually involves not just the provision of information about performance, but also the possibility of debate, of questions by the forum and answers by the actor, and eventually of judgment of the actor by the forum." Ingrid Opdebeek and Stephanie De Somer [63] develop a similar argument in the context of administrations.

### 3.5 Contestability ensures a balanced and mutually beneficial collaboration between ADS and humans

In addition to being essential components of accountability and legitimacy, justifications and contestations have the potential to bring a radical change to the integration of ADS within human environments. As firmly asserted by Daniel N. Kluttz and his co-authors [47], "Contestability fosters engagement rather than passivity, questioning rather than acquiescence. [...] Contestability can support critical, generative, and responsible engagement between users and algorithms, users and system designers, and ideally between users and those subject to decisions (when they are not the users), as well as the public." In that sense, contestability contributes to preserve the autonomy of the human decision maker. Indeed, even if the human decision maker does not have any formal obligation to follow the suggestion of the ADS, his/her autonomy is questionable if he/she does not have any possibility to contest it. This has led Daniel N. Kluttz and his co-authors to call for an evolution of the regulation to emphasize contestability rather than explainability: "Regulatory approaches should seek to put professionals and decision support systems in conversation, not position professionals as passive recipients of system wisdom who must rely on out-of-system mechanisms to challenge them. For these reasons, calls for explainability fall short and should be replaced by regulatory approaches that drive contestable design."

### 3.6 When should justifiability and contestability be required ?

It should be clear, however, that justifiability and contestability cannot be imposed to any ADS in any context. Therefore, the next question to be addressed

[19] Reuben Binns' example [12] illustrates the fact that justifications and contestations are essential parts of accountability : "For instance, a bank deploying an automated credit scoring system might be held accountable by a customer whose loan application has been automatically denied. Accountability in this scenario might consist of a demand by the customer that the bank provide justification for the decision; [...] and a final step, in which the customer either accepts the justification, or rejects it, in which case the bank might have to revise or reprocess their decision with a human agent, or face some form of sanction."

is the identification of the conditions in which they should be required. The first criterion to take into account is the potential impacts of the decisions on individuals and on society in general. In this regard, the conditions set out in the GDPR are clearly insufficient as they refer to persons "subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." They are too restrictive for at least two reasons: they seem to exclude decisions taken on the basis of an ADS (but not "based only" on it) and decisions that would not have a significant impact on individuals. For example, an ADS used to suggest shopping recommendations, to target political ads, or to select route choices can have a strong impact on economic players or on society without significantly affecting any particular individual. Therefore, the general rule should be that an ADS must be accountable, and therefore justifiable and contestable, as soon as it can have an impact on any stakeholder (citizen, professional, economic operator, etc.), any social group or on society as a whole (social impact, political impact, etc.).

The next, and more complex questions about this accountability requirement are : accountable towards whom and how ? The operationalization of justifiability and contestability are discussed in the next section. As far as the beneficiary is concerned, the answer depends primarily on the type of system and its mode of interaction with its environment. Let us consider, at one end of the spectrum, decision components embedded into critical systems such as automatic metro control systems or a car braking systems. This type of components can undoubtedly be considered as ADS that can have a major impact on human lives. However, they make decisions automatically without any interaction with any individual. No metro passenger or car driver would obviously require justifications about such components, which does not mean that they should not be subject to accountability requirements. For this type of ADS, the beneficiary of the justifications and the potential issuer of contestations could be an auditor or an expert working for a certification body.

Another type of system to be considered are the ADS used by professionals (bank officers, lawyers, judges, clinicians, etc.) to obtain information (predictions, recommendations, diagnosis, etc.) falling within their area of expertise. In this context, it is critical that the professional gets appropriate justifications and has means to contest the outcomes of the ADS. It is a key condition to ensure that the ADS is used appropriately and to foster "engagement rather than passivity, questioning rather than acquiescence" [47]. This type of justification should come in addition to accountability requirements towards auditors or certification bodies in critical sectors such as health-care, justice and banking. It should also facilitate the accountability of the professional himself/herself towards the people affected by the decisions (patient, litigant, customer, etc.).

At the other end of the spectrum, some ADS are involved in everyday services such as recommendation systems, sometimes even without the knowledge of the users (targeted ads, information feeds, etc.). Like other types of critical systems, such ADS should be accountable towards independent auditors (for

example regulation authorities) to ensure that they are not misleading and they do not lead to unfair practices, discriminations or manipulations. In addition, justifications and possibilities of contestation should be provided to lay users in a very simple and accessible way (for example, by refusing certain types of ads). In fact, these obligations may become legal requirements if the Digital Services Act (DSA) proposed by the European Commission in December 2020 [30] is adopted as currently drafted. In its summary of its consultations, the DSA observes that "several stakeholders, in particular civil society and academics, pointed out the need for algorithmic accountability and transparency audits, especially with regard to how information is prioritized and targeted. Similarly, regarding online advertising, stakeholder views echoed the broad concerns around the lack of user empowerment and lack of meaningful oversight and enforcement." Needless to say, justifications and contestations must take very different forms depending on the beneficiary. We present a range of technical solutions to implement them in the next section.

## 4 Operationalization of justifiability and contestability

Different types of justifications and different modes of contestation are appropriate depending on the level of expertise of the stakeholders and the nature of their involvement in the ADS. We can distinguish four main types of situations:

1. Stakeholders involved in an ex-ante algorithmic impact assessment of the ADS.
2. Experts who intervene either upstream (e.g. certification bodies) or downstream (e.g. control authorities or auditors).
3. Professionals involved in the operational phase of the ADS (for example doctors, bank officers or judges) and making decisions potentially based on its outcomes. No assumption can be made about the level of expertise of these professionals regarding algorithmic techniques.
4. Laypersons affected by the outcomes of the ADS[20].

Each type of situation calls for specific types of justifications and contestations which can be supported by different combinations of organizational and technical instruments. The first case is different in nature since it occurs before the start of the design of the system. Furthermore, algorithmic impact assessments constitute a specific task in themselves; therefore, we postpone their discussion until Section 6 and present successively the support to be provided to experts in Section 4.1 and to non experts (professionals and laypersons) in Section 4.2.

### 4.1 Justifications and contestations for experts

The justifications to be provided to experts should include all the documents typically expected in a certification or audit process: requirements, documentation

[20] Because they use it, explicitly or implicitly, or they are subject to decisions taken by professionals.

about the design and the implementation of the ADS, testing procedure, test results, execution logs, etc. In addition, any ADS relying on machine learning should be accompanied with details about the learning dataset (the conditions of its creation, the collection, cleaning, labeling processes, etc.) and the model (intended use, relevant factors, performances, etc.). In order to facilitate communication and to ensure comprehensiveness, the information could be provided in standardized formats such as *Datasheets for Data Sets* [34] and *Model Cards* [57]. Justifications are complete only if they establish a continuous link between the high-level objectives of the ADS (the applicable norms, for example non-discrimination, reduction of recidivism rate, or compliance with a given legal requirement) and its implementation, with arguments about the fact that the implementation complies with the objectives. These arguments can take different forms, which can be more or less conclusive, from high-level testing, to detailed and systematic testing or even formal proofs in certain situations[21]. For example, certain high-level certification schemes require or favour formal (mathematical) proofs of consistency between different levels of refinement of the system. Such proofs have been provided for critical code embedded in transport systems [7] and smart cards [22], for example. In general, compliance is established through several levels of refinements including at least the requirements (applicable norms), the specification (characterization of the means used to meet the requirements) and the code itself (implementation of the system). For instance, in the case of an automatic metro, the most important norm should be to protect the lives of the passengers; the specification should include constraints on the operation of the trains (speed, acceleration, deceleration, distance between two trains, etc.) and the code should comply with these constraints. It should be clear, however, that formal proofs are only possible (and required) in extreme situations where the specifications of the system (or the essential properties that it must satisfy) can be defined in a precise way and the implementation itself is amenable to automatic or semi-automatic verification. These conditions are not met for most ADS relying on machine learning since their specifications can generally not be characterized by logical formulas[22] and their performances are better assessed through statistical means. A lot of work has been done in the computer science community to address this issue, in particular to define and evaluate different metrics of fairness and accuracy [19].

As discussed in the previous section, contestations and justifications are dual. Therefore, experts can use the same means to challenge the justifications provided by the operator (further testing, verification of further properties, etc.).

[21] As stated by Finale Doshi-Velez and Been Kim [28], "for complex tasks, the end-to-end system is almost never completely testable; one cannot create a complete list of scenarios in which the system may fail. Enumerating all possible outputs given all possible inputs be computationally or logistically infeasible, and we may be unable to flag all undesirable outputs."

[22] For example, as mentioned in [28], "the human may want to guard against certain kinds of discrimination, and their notion of fairness may be too abstract to be completely encoded into the system"

### 4.2 Justifications and contestations for professionals and laypersons

In contrast to experts, professionals and affected people are involved in the operational phase of the ADS and they have to cope with particular situations (their own situation for affected people; the situation of the persons they have to serve - patients, litigants, or customers - for professionals). Therefore, they have a different focus (more "local" than "global") and they require different modalities for justifications and contestations (more interactive and less demanding in terms of expertise). This type of interaction should be supported by appropriate tools. However, little work has been done so far to facilitate justifications and contestations for non experts. One of the rare examples is the Algocate tool [37, 38], which can be illustrated on the fictive example of a national vaccine allocation system. The inputs of this ADS are medical data (previous diseases, risk factors, etc.) and demographic data (age, region, profession, etc.). The outcomes are individual vaccine attribution decisions.

1. Regional equality rule: regions should receive a number of vaccine doses in proportion to the size of their population.
2. Death limitation: vaccine allocation should minimize the global number of deaths in the country (e.g. over a three-month time span).
3. Spread limitation: vaccine allocation should minimize the global spreading of the disease in the country (e.g. over a three-month time span).

**Fig. 1.** Examples of norms for a vaccine allocation ADS

Such allocation systems give rise to various ethical considerations [64]. For the sake of our example, we consider a simplified situation with three principles elicited as norms in Fig. 1. These norms have different sources of legitimacy (based on fairness or utilitarian considerations) and can be in tension. For example, even if death and spread limitations seem to support each other in the long term, they result in different vaccination strategies. The first one prioritizes people with high death risk (for instance the elderly and people with co-morbidities) while the second one prioritizes people with a high potential of spreading the virus (e.g. medical staff, teachers and professionals with many contacts). Norms can be hierarchical. In this example, we assume that the first norm, which is subject to a decree, prevails over the two other norms (which have the same priority). Technically speaking, Algocate includes three types of norms (rule, objective and reference) [38]. The first norm in Fig. 1 is an example of rule while the two others are examples of objectives.

Algocate does not take for granted that the legitimacy of a norm is accepted by all parties. For instance, an utilitarian may challenge the first norm of Fig. 1 while a Kantian may consider that the second norm should prevail. In such situations, if a party refers to a norm or hierarchy that is not accepted by the

other party, the contestation-justification protocol requires the intervention of the human decision maker. The benefit of the protocol in such cases is that a party relying on a norm is compelled to elicit it and to submit it to the approval of the other party (or the human decision maker as a last resort). Contestations can take different forms depending on the status of the disputed norm (e.g. legal proceedings for a legal norm).

U1 I challenge the decision because I am older than 60 and I am practicing as nurse, so I think that I should receive a vaccine.
A1 First, candidates (as you) who are older than 60 and who practice as nurses have an average spreading factor of 2.3, which is significantly above the average of 1.2.
A2 However, providing you a vaccine would breach the regional equality decree.

**Fig. 2.** First example of interaction with *Algocate*

Fig. 2 presents an example of interaction of a user with Algocate in which statements are expressed in a restricted natural language[23]. The statements of the user are labelled with *Ui* and Algocate answers are labelled with *Ai*:

- In *U1*, the user challenges the rejection of his/her vaccine application and expresses the reasons why he/she believes that his/her application should be accepted.
- In *A1*, Algocate provides a piece of evidence to support the user's challenge based on the spread limitation norm. This evidence is generated from a spread factor database, which has been built using a spread risk estimation model.
- In *A2*, Algocate provides a justification of the decision based on the regional equality rule.
- Because the regional equality rule prevails, the interaction may stop here as Algocate has provided evidence which, according to its settings, is sufficient to justify the decision. The contestation may still proceed but it should then take the form of legal proceedings to challenge the regional equality decree.

Fig. 3 presents another example of interaction with Algocate:

- In *U1*, the user challenges the rejection of his/her vaccine application.
- In *A1*, Algocate provides evidence of non application of the regional equality rule.
- In *A2*, Algocate provides evidence based on the death limitation norm to support the user's challenge. The evidence is a refinement of the challenge including an additional attribute, disease D. The evidence is generated from a death risk database, which has been built with a death risk estimation model.

[23] With standard keywords, such as first, furthermore, however, etc.

U1 I challenge the decision because I am older than 70, so I think that I should receive the vaccine.
A1 First, you belong to region R which has not yet reached its quota of vaccine.
A2 Furthermore, candidates (as you) who are older than 70 and who have disease D have a death risk of 5 %, which is significantly above the average of 0.1 %.
A3 However, candidates (as you) who are older than 70 and who do not practice in the medical sector have a an average spreading factor of 0.8, which is significantly below the average of 1.2.

**Fig. 3.** Second example of interaction with *Algocate*

- In *A3*, Algocate provides a justification for the rejection decision based on the spread limitation norm.
- The justification shows that the system has prioritized the spread limitation objective over the death limitation objective. If he/she thinks he/she has other ways to contest the decision, the user may initiate another iteration step of Algocate with another challenge. Otherwise, he/she must either accept the justification or contest the legitimacy of the prevalence of spread limitation norm.

The interactions presented in Fig. 2 and Fig. 3 show that Algocate is useful both for users contesting the decisions of the ADS (e.g. individuals affected by the decisions or professionals who are not sure about the suggestion of the ADS) and for users who want to justify them (usually the operator of the ADS). Algocate is neutral, in the sense that it is designed to find the best arguments (i.e. statements with evidence supporting them) for both parties. This neutrality is of prime importance, given the usual imbalance of powers between individuals who are affected by the decisions and the designers or operators of the ADS. As stated by Reuben Binns [12], "Giving absolute primacy to either the decision-maker or the decision-subject would render algorithmic accountability too one-sided, allowing one party to hold the other to standards that they could not reasonably accept."

Algocate shows that justifiability and contestability are possible also for non-experts, provided that they are supported by appropriate tools. Another project going in the same direction is described in [41]: the paper acknowledges the need "to provide mechanisms for users to challenge model predictions" and states that "doing so may require users to marshal evidence and create counter narrative that argue precisely why they disagree with a conclusion drawn by an AI system." However, it does not describe precise mechanisms to support this type of contestation.

## 5 Related work

In the field of explainable AI, the distinction between explanations and justifications is sometimes blurred. For example, [14] refers to explanations "which are presumed to be justifying" and [50] considers that justifications are ways to make understandable the inner operations of a complex system (in a white-box setting). However, a series of works [13, 15, 16, 55, 61, 69], refer to justifications as ways of ensuring that a decision is good (in contrast to understanding a decision), which is in line with the definition proposed in this paper. Also, even if the word "justification" is not used in the paper, [25] introduces an "explicability" principle for AI taken "both in the epistemological sense of 'intelligibility' and in the ethical sense of 'accountability'". The normative nature of justifications was also mentioned in the field of intelligent systems [48]: "an intelligent system exhibits justified agency if it follows society's norms and explains its activities in those terms." However, these norms are not characterized precisely in [48]. On this matter, [49] qualifies explanations as "unjustified" when there are not supported by training data. Therefore, justifiability applies to explanations in this context rather than to the decisions themselves. From a different perspective, [24], distinguishes different types of justifications but they concern only the performances of inductive machine learning techniques.

The term "justification" is sometimes used in the field of autonomous agents to refer to motivations to perform a specific action. In this context, norms are made available to autonomous agents to make their decisions. For example, the framework introduced in [51] relies on norms (called "principles") inferred from ethical judgments using inductive logic programming. In a nutshell, a value-driven agent refers to an ethical preference ordering of the actions before making any decision. It is worth noting that it is appropriate to present justifications as a type of explanation in this setting since norms are in internalised in agents.

The interest for more interactions with humans in the design and exploitation phases of machine learning systems takes different forms [1]. The need to conceive explanations as an interactive process has been argued by several authors [56, 58]. The "human-in-the-loop" approach leverages on human feedback during the training process to obtain more accurate classifiers [46]. A lot of work has also been done on argumentation and dialog games [6, 11, 51, 53] but the focus in these areas is generally the logical structure of the framework to express and to relate arguments or the protocol to exchange arguments. Closer to the notion of justification, [43] relies on "debates" between two competing algorithms exchanging arguments and counterarguments to convince a human user that their classification is correct. However, the goal of this work is to "align an agent's actions with the values and preferences of humans" which is seen as a "training-time problem".

The limitations of transparency (seen as "looking inside the black box") for accountability has already been argued in [4]. In the area of administration, [63] concludes that "the process of giving a proper justification for an administrative decision involves more than just transparency for the sake of transparency. It is an argumentative process that provides targeted information." The need to

provide appropriate justifications of judicial decisions has also been analyzed by several authors [26, 29, 73][24]. For example, [26] refers to justifications "as a powerful preventive of wrong decisions to encourage uniformity across decision-making bodies and to make decisions somewhat more acceptable to a losing claimant."

The advantages of justifications over explanations (in the sense used in this paper) to enhance trust in ADS has also been analyzed by Karl de Fine Licht and Jenny de Fine Licht [32] who argue that "a limited form of transparency that focuses on providing justifications for decisions has the potential to provide sufficient grounds for perceived legitimacy in AI decision-making."

As far as operationalization is concerned, the notion of "design transparency" proposed by Michele Loi and his co-authors [52] is close to what we described as "justifications for experts" in Section 4.1. Design transparency "requires giving information on various elements: the goal that the algorithm pursues, the mathematical constructs into which the goal or its proxy is translated in order to be implemented in the algorithm, and the tests and the data with which the performance of the algorithm was verified." The authors coin the expressions "value transparency", "translation transparency" and "performance transparency" to denote these three elements.

## 6 Conclusion

AI and ADS can bring great benefits to society but there is strong risk of public backlash if ethical issues are not considered seriously [10, 21, 32, 62][25]. As firmly asserted by Mariarosaria Taddeo and Luciano Floridi [70], "Ethical regulation of the design and use of AI is a complex but necessary task. The alternative may lead to devaluation of individual rights and social values, rejection of AI-based innovation, and ultimately a missed opportunity to use AI to improve individual well-being and social welfare." In this paper, we have argued that justifiability and contestability are key conditions to ensure the legitimacy of ADS that can have a significant impact on individuals' lives or on society in general. In this regard, it is worth noting that the proposal for a regulation "on a Single Market For Digital Services (Digital Services Act)" [30] published by the European Commission in December 2020 also emphasises the ability for users of online platforms to contest decisions. For example, Recital 44 states that "Recipients of the service should be able to easily and effectively contest certain decisions of online platforms that negatively affect them. Therefore, online platforms should be required to provide for internal complaint-handling systems, which meet certain conditions aimed at ensuring that the systems are easily accessible and lead to swift and fair outcomes."

[24] Note that the word "explanation" is used in the sense of "justification" by the authors of [29], or "motivation" in legal parlance.

[25] As an illustration, a recent survey [10] conducted by BEUC across nine EU countries shows that in all of them, a majority of people "agree or strongly agree that companies are using AI to manipulate consumer decisions."

In this paper, we have also shown that justifiability and contestability are not illusory objectives: they can be operationalized in different ways for different types of recipients. Even though we have focused on technical means in this paper, we do not want to suggest that justifiability and contestability could be reduced to technical issues. The role of the technical means presented here is to support an overall combination of legal, political and organizational framework to enhance the legitimacy of ADS. Also, it should be clear that the deployment of any critical ADS should be preceded by a multistakeholder algorithmic impact assessment [19, 20, 44, 60, 65]. The means presented in Section 4 (in particular, in Section 4.1) and the mere elicitation of applicable norms (as argued in Section 3.3) can provide useful inputs to the deliberation about the legitimacy of the ADS but they should obviously not overshadow the debate, which must involve subjective (political, philosophical, ethical) considerations. As stated by Richard Berk and his co-authors [8], " these are matters of values and law, and ultimately, the political process. They are not matters of science." In return, the algorithmic impact assessment should provide guidance about the level of requirements to be imposed to the ADS, in particular in terms of justifiability and contestability. A good illustration of this approach is the Canadian Directive on Automated Decision-Making [35] which introduces four "Impact Assessment Levels" associated with different "Impact Level Requirements". Even though it applies only to administrative decisions of the federal government, it shows the way to the integration of algorithmic impact assessment into the development cycle of all critical ADS. In Europe, the recent recommendations of the European Parliament [31] points in the same direction, considering "that the determination of whether artificial intelligence, robotics and related technologies should be considered high-risk, and thus subject to mandatory compliance with legal obligations and ethical principles as set out in the regulatory framework for AI, should always follow from an impartial, regulated and external ex-ante assessment based on concrete and defined criteria."

## References


1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. p. 1–18. ACM Press (2018). https://doi.org/10.1145/3173574.3174156, http://dl.acm.org/citation.cfm?doid=3173574.3174156
2. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
3. Alvarez-Melis, D., Jaakkola, T.S.: A causal framework for explaining the predictions of black-box sequence-to-sequence models (2017)
4. Ananny, M., Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society **20**(3), 973–989 (2018). https://doi.org/10.1177/1461444816676645, https://doi.org/10.1177/1461444816676645
5. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. arXiv:1910.10045 [cs] (Dec 2019), http://arxiv.org/abs/1910.10045, arXiv: 1910.10045
6. Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G., Thimm, M., Villata, S.: Towards artificial argumentation. AI Magazine **38**(3), 25–36 (Oct 2017). https://doi.org/10.1609/aimag.v38i3.2704, https://www.aaai.org/ojs/index.php/aimagazine/article/view/2704
7. Beek, M.H., Gnesi, S., Knapp, A.: Formal methods for transport systems. Int. J. Softw. Tools Technol. Transf. **20**(3), 237–241 (Jun 2018). https://doi.org/10.1007/s10009-018-0487-4, https://doi.org/10.1007/s10009-018-0487-4
8. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art (2017)
9. Bernstein, S.: Legitimacy in global environmental governance. International Journal of Comparative Labour Law and Industrial Relations **1**, 139–166 (2005)
10. BEUC The European Consumer Organization: Artificial Intelligence: what consumers say. Findings and policy recommendations of a multi-country survey on AI. (2019)
11. Bex, F., Walton, D.: Combining explanation and argumentation in dialogue. Argument and Computation **7**(1), 55–68 (2011)
12. Binns, R.: Algorithmic accountability and public reason. Philosophy and Technology **31** (2018), https://doi.org/10.1007/s13347-017-0263-5
13. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI-17 Workshop on Explainable AI (XAI). p. 8 (2017)
14. Biran, O., McKeown, K.: Human-centric justification of machine learning predictions. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 1461–1467 (2017). https://doi.org/10.24963/ijcai.2017/202, https://doi.org/10.24963/ijcai.2017/202
15. Biran, O., McKeown, K.R.: Justification narratives for individual classifications. In: ICML (2014)

16. Biran, O., McKeown, K.R.: Human-centric justification of machine learning predictions. In: IJCAI. p. 1461–1467 (2017)
17. Black, J.: Constructing and contesting legitimacy and accountability in polycentric regulatory regimes. Regulation & Governance **2**(2), 137–164 (2008). https://doi.org/https://doi.org/10.1111/j.1748-5991.2008.00034.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1748-5991.2008.00034.x
18. Bovens, M.: Analysing and Assessing Public Accountability. A Conceptual Framework. European Governance Papers (EUROGOV) 1, CONNEX and EUROGOV networks (Jan 2006), https://ideas.repec.org/p/erp/eurogo/p0005.html
19. Castelluccia, C., Le Métayer, D.: Understanding algorithmic decision-making: Opportunities and challenges. Report for the European Parliament (Panel for the Future of Science and Technology - STOA) (2019), https://www.europarl.europa.eu/stoa/en/document/EPRS STU(2019)624261
20. Castelluccia, C., Le Métayer, D.: Position paper: Analyzing the impacts of facial recognition. In: Antunes, L., Naldi, M., Italiano, G.F., Rannenberg, K., Drogkaris, P. (eds.) Privacy Technologies and Policy - 8th Annual Privacy Forum, APF 2020, Lisbon, Portugal, October 22-23, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12121, pp. 43–57. Springer (2020). https://doi.org/10.1007/978-3-030-55196-4_3, https://doi.org/10.1007/978-3-030-55196-4_3
21. Center for Data Ethics and Innovation (CDEI): AI Barometer Report (2020)
22. Chetali, B., Nguyen, Q.H.: Industrial use of formal methods for a high-level security evaluation. In: Cuellar, J., Maibaum, T., Sere, K. (eds.) FM 2008: Formal Methods. pp. 198–213. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
23. Christin, A., Rosenblat, A., d. boyd: Courts and predictive algorithms. In: Primer for the Data and Civil Rights Conference: A New Era of Policing and Justice (2015)
24. Corfield, D.: Varieties of justification in machine learning. Minds and Machines **20**(2), 291–301 (Jul 2010). https://doi.org/10.1007/s11023-010-9191-1
25. Cowls, J., Floridi, L.: Prolegomena to a white paper on an ethical framework for a good ai society. SSRN Electronic Journal (01 2018). https://doi.org/10.2139/ssrn.3198732
26. Crawford, K., Schultz, J.: Big data and due process: Toward a framework to redress predictive privacy harms. Boston College Law Review **55** (2014), https://lawdigitalcommons.bc.edu/bclr/vol55/iss1/4
27. Danaher, J.: The threat of algocracy: Reality, resistance and accommodation. Philosophy & Technology **29**, 245–268 (2016)
28. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
29. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., Wood, A.: Accountability of ai under the law: The role of explanation (2019)
30. European Commission: Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Tech. rep., EC (2020)
31. European Parliament: Report with recommendations to the commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020)
32. de Fine Licht, K., de Fine Licht, J.: Artificial intelligence, transparency, and public decision-making. AI & Society (35), 917—-926 (2020), https://doi.org/10.1007/s00146-020-00960-w

33. Fredriksson, M., Tritter, J.: Disentangling patient and public involvement in healthcare decisions: why the difference matters. Sociology of health & illness **39 1**, 95–111 (2017)
34. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., au2, H.D.I., Crawford, K.: Datasheets for datasets (2020)
35. Government of Canada: Directive on automated decision-making (2019)
36. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) **51**(5), 93 (2018)
37. Henin, C., Le Métayer, D.: Towards a framework for challenging ml-based decisions. In: DeceptECAI 2020 - 1st International Workshop on Deceptive AI (2020)
38. Henin, C., Le Métayer, D.: A Framework to Contest and Justify Algorithmic Decisions (Feb 2021), https://hal.inria.fr/hal-03127932, working paper or preprint
39. Henin, C., Le Métayer, D.: A generic framework for black-box explanations. In: Proceedings of the International Workshop on Fair and Interpretable Learning Algorithms (FILA 2020). IEEE (2020)
40. Hildebrandt, M.: Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. Theoretical Inquiries in Law **20**, 83–122 (2019)
41. Hirsch, T., Merced, K., Narayanan, S., Imel, Z.E., Atkins, D.C.: Designing contestability: Interaction design, machine learning, and mental health. In: Proceedings of the 2017 Conference on Designing Interactive Systems. p. 95–99. DIS '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3064663.3064703, https://doi.org/10.1145/3064663.3064703
42. HLEG, A.: Ethics guidelines for trustworthy AI. Tech. rep., European Commission High-Level Expert Group on Artificial Intelligence (2019)
43. Irving, G., Christiano, P., Amodei, D.: AI safety via debate. arXiv:1805.00899 [cs, stat] (Oct 2018), http://arxiv.org/abs/1805.00899, arXiv: 1805.00899
44. Kaminski, M.E., Malgieri, G.: Algorithmic impact assessments under the GDPR: producing multi-layered explanations. International Data Privacy Law (12 2020). https://doi.org/10.1093/idpl/ipaa020, https://doi.org/10.1093/idpl/ipaa020, ipaa020
45. van Kersbergen, K., van Waarden, F.: 'governance' as a bridge between disciplines. cross-disciplinary inspiration regarding shifts in governance and problems of governability, accountability, and legitimacy. European Journal of Political Research **43**, 143 – 171 (2004)
46. Kim, B.: Interactive and interpretable machine learning models for human machine collaboration p. 143
47. Kluttz, D.N., Kohli, N., Mulligan, D.K.: Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions, p. 137–152. Cambridge University Press (2020)
48. Langley, P.: Explainable, normative, and justified agency. Proceedings of the AAAI Conference on Artificial Intelligence **33**, 9775–9779 (Jul 2019). https://doi.org/10.1609/aaai.v33i01.33019775
49. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: Unjustified counterfactual explanations. arXiv:1907.09294 [cs, stat] (Jul 2019), http://arxiv.org/abs/1907.09294, arXiv: 1907.09294
50. Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing neural predictions. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Lan-

guage Processing. p. 107–117. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/D16-1011, http://aclweb.org/anthology/D16-1011
51. Liao, B., Anderson, M., Anderson, S.L.: Representation, justification, and explanation in a value-driven agent: an argumentation-based approach. AI and Ethics pp. s43681–020–00001–00008 (Sep 2020). https://doi.org/10.1007/s43681-020-00001-8
52. Loi, M., Ferrario, A., Vigano, E.: Transparency as design publicity: explaining and justifying inscrutable algorithms. Ethics and Information Technology (2020), https://doi.org/10.1007/s10676-020-09564-w
53. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: A grounded interaction protocol for explainable artificial intelligence. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. p. 1033–1041. AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2019)
54. Malgieri, G., Comandé, G.: Why a right to legibility of automated decision-making exists in the general data protection regulation. International Data Privacy Law (2017)
55. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267** (2017). https://doi.org/10.1016/j.artint.2018.07.007
56. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: Beware of inmates running the asylum. In: IJCAI-17 Workshop on Explainable AI (XAI). vol. 36 (2017)
57. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19 (2019). https://doi.org/10.1145/3287560.3287596, http://dx.doi.org/10.1145/3287560.3287596
58. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. arXiv:1811.01439 [cs] (Nov 2018). https://doi.org/10.1145/3287560.3287574, http://arxiv.org/abs/1811.01439, arXiv: 1811.01439
59. Monahan, J., Skeem, J.: Risk assessment in criminal sentencing. Annual review of clinical psychology **12**, 489–513 (2016)
60. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. Science and engineering ethics **26**(4), 2141–2168 (2020)
61. Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A.: Explanation in human-AI systems: A literature meta-review synopsis of key ideas and publications and bibliography for explainable AI p. 204 (2019)
62. Narayanan, A.: How to recognize AI snake oil (2019), https://www.cs.princeton.edu/ arvindn/talks/MIT-STS-AI-snakeoil.pdf
63. Opdebeek, I., Somer, S.D.: The duty to give reasons in the european legal area: a mechanism for transparent and accountable administrative decision-making? a comparison of belgian, dutch, french and eu administrative law. Rocznik Administracji Publicznej **2** (2016)
64. Persad, G., Wertheimer, A., Emanuel, E.J.: Principles for allocation of scarce medical interventions. The Lancet **373**(9661), 423–431 (Jan 2009). https://doi.org/10.1016/S0140-6736(09)60137-9
65. Reisman, D., Schultz, J., Crawford, K., Whittaker, M.: Algorithm Impact Assessment: A Practical Frameworks for Public Agency Accountability (AINow Institute Report) (2018)
66. Rouvroy, A.: The end(s) of critique: data-behaviourism vs. due process. In: Privacy, Due Process and the Computational Turn. Philosophers of Law Meet Philosophers of Technology. Routledge (2013)

67. Rouvroy, A.: A few thoughts in preparation for the discrimination and big data conference organized by Constant at the CPDP (2015)
68. Suchman, M.C.: Managing legitimacy: Strategic and institutional approaches. The Academy of Management Review **20**(3), 571–610 (1995), http://www.jstor.org/stable/258788
69. Swartout, W.R., Swartout, W.R.: Producing Explanations and Justifications of Expert Consulting Programs (1981)
70. Taddeo, M., Floridi, L.: How ai can be a force for good. Science **361**(6404), 751–752 (2018). https://doi.org/10.1126/science.aat5991, https://science.sciencemag.org/content/361/6404/751
71. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation (12 2016). https://doi.org/10.1093/idpl/ipx005
72. Waldman, A.E.: Power, process, and automated decision-making. Fordham Law Review (88) (2019), https://ssrn.com/abstract=3461238
73. Wroblewski, J.: Legal decision and its justification. Logique et Analyse **14**(53/54), 409–419 (1971), http://www.jstor.org/stable/44074477