TU Delft

Delft University of Technology

Transparency for AI systems

A value-based approach

Buijsman, Stefan

DOI
10.1007/s10676-024-09770-w

Publication date
2024

Document Version
Final published version

Published in
Ethics and Information Technology

Citation (APA)
Buijsman, S. (2024). Transparency for AI systems: A value-based approach. *Ethics and Information Technology*, *26*, Article 34. https://doi.org/10.1007/s10676-024-09770-w

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Ethics and Information Technology (2024) 26:34
https://doi.org/10.1007/s10676-024-09770-w

ORIGINAL PAPER

Transparency for AI systems: a value-based approach

Stefan Buijsman[1]

Accepted: 25 April 2024
© The Author(s) 2024

Abstract
With the widespread use of artificial intelligence, it becomes crucial to provide information about these systems and how they are used. Governments aim to disclose their use of algorithms to establish legitimacy and the EU AI Act mandates forms of transparency for all high-risk and limited-risk systems. Yet, what should the standards for transparency be? What information is needed to show to a wide public that a certain system can be used legitimately and responsibly? I argue that process-based approaches fail to satisfy, as knowledge about the development process is insufficient to predict the properties of the resulting system. Current outcome-based approaches [Mitchell et al., 2019; Loi et al., 2021] are also criticized for a lack of attention to the broader socio-technical system and failure to account for empirical results that show that people care about more than just the outcomes of a process [as reported by Meyerson et al. (Procedural justice and relational theory: Empirical, philosophical, and legal perspectives, Taylor & Francis, 2021)]. Instead, I propose value-based transparency, on which the information we need to provide is what values have been considered in the design and how successful these have been realized in the final system. This can handle the objections to other frameworks, matches with current best practices on the design of responsible AI and provides the public with information on the crucial aspects of a system's design.

Keywords Transparency · Legitimacy · AI ethics · Design for values

Introduction

What kind of information should be provided about algorithms (perhaps specifically AI systems) to the general public? This is a pressing question now that governments and companies are using algorithms more and more, with potentially massive impacts such as large-scale discrimination in decision making [see Amnesty International (2021) for an example]. As a result, there is a growing demand for transparency about the use of such systems, not least as one of the basic requirements in the EU AI Act,[1] which in the current draft requires forms of transparency for both high-risk and limited risk systems (Varošanec, 2022). This should help make users aware that they are interacting with AI systems, as well as show (especially for the high-risk systems) that it is responsible to use these systems. As such, it is a different kind of transparency than the one at stake in discussions of the explainability of AI systems (Das & Rad, 2020) and in the XAI literature. Explainability focuses on the reasons for singular decisions/outputs, and thus on making the model more interpretable. In contrast, my use of transparency here refers to the information provided about the system as a whole. There is a crucial difference between the two: transparency focuses on the AI system as a whole, and aims to establish that the system (with all the relevant features) may be used. Explainability, on the other hand, is best seen as a feature of the (socio-)technical AI system that aims to support decision making with the model. So transparency encompasses (among others) information on the accuracy and fairness of the system and the effectiveness of human oversight. Explainability has as goal to clarify the reasons for particular outputs of the AI system, for example through a generalization that shows how the output of the system depends on the inputs (Buijsman, 2022). The question I am concerned with here is then: what information about the socio-technical system as a whole needs to be provided? How explainable a system is may then be part of that

✉ Stefan Buijsman
s.n.r.buijsman@tudelft.nl

[1] TU Delft, Jaffalaan 5, 2628 BX Delft, The Netherlands

[1] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending certain Union Legislative Acts COM (2021) 206 final (and encompassing Annexes 1 to 9) [AI Act proposal], https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN.

Published online: 15 May 2024

required information about the socio-technical system. Vice versa, individual decisions made with the system (and what information needs to be provided then) are not considered here, but only the general functioning of the system.

This system-level information is highly relevant to establish the legitimacy of the use of such systems. The public sector needs to show the (political) legitimacy of the AI systems in use, as there is a need to demonstrate that decisions affecting citizens are made in an acceptable manner. Transparency should allow for the establishment of this legitimacy, also by offering opportunities for e.g. investigative journalists to critically reflect on the use of algorithms. But for the private sector, too, we can wonder about the acceptability of the algorithms they use. Do they adhere to the rules set out to maintain their license to do business? And is, for example, the way they evaluate candidates when using AI to conduct job interviews legitimate? Companies may not have the same pressure to make this information public as governments, but still need to (a) decide whether a system they procure from a supplier meets their demands and (b) be able to show regulators that AI systems that they deploy meet legal requirements. For both goals a framework for transparency is needed. So, what information should be provided with these goals for transparency in mind? I discuss this question first theoretically, based on different views of legitimacy, in Sect. "What kind of transparency do we need?". From there, I move on to different (current) approaches to transparency regarding AI systems.

Governments are starting to formulate their own approaches in the form of algorithm registers, such as the Dutch Algorithm Register, http://algoritmes.overheid.nl. These set standards for the amount and kind of information that is given about algorithms, and have so far opted for a primarily process-based approach as I will discuss in Sect. "Current approaches: Process-based transparency". At the same time, there are academic accounts of what transparency should consist of. There are the more practically oriented datasheets and model cards (Gebru et al., 2021; Mitchell et al., 2019) as well as the more philosophical account called design publicity (Loi et al., 2021). These tend to be more outcome-based, focusing on how well an algorithm fulfills its goal and how this was established. I'll discuss them in Sect. "Current approaches: Outcome-based transparency". I will argue that both approaches to transparency have shortcomings: process-based accounts fail to show whether the system that results from the described development process meets (ethical) standards because the behaviour of (machine-learning) systems is hard to predict based purely on the steps taken by developers. On the other hand, current outcome-based accounts take too narrow a view of AI systems and miss parts of the process that have been shown to be crucial to perceptions of legitimacy. Instead, I'll argue in Sect. "Value-based transparency" that an approach based on the whole range of values we want to realise in an AI system, effectively extending current outcome-based approaches, manages to handle these issues nicely and fits with current best practices on the realisation of responsible AI.

What kind of transparency do we need?

I [and others, such as Loi et al. (2021)] take establishing the legitimacy of AI systems as one of the central goals of transparency. We need to have the right information about these AI systems to determine if their use is acceptable, and currently this is often missing. If we consider the different sources of political legitimacy in the philosophical literature (Peter, 2017), then these help to pinpoint what might be needed to meet that goal. To start with, there are theories of legitimacy that consider consent to be central to legitimacy in one way or another (Estlund, 2009; Greene, 2016; Simmons, 2001). These present the question whether (ideal) agents have enough information to consent to the use of these algorithms.

The first question here is what people would consent to exactly. There are, at least, two options on how broad we construe consent: either we consent to a specific AI system without regards to the broader decision making process of which it is a part, or we consent to the entire decision making process (AI included). I am of the opinion that it is the broader socio-technical system and thus the entire decision-making process, that is the right unit for consent. Consent, after all, should focus on how decisions are made, and whether we find the way that we are treated acceptable. As AI outputs typically do not translate directly into decisions and the rest of the process has serious impact on the decision making we should therefore focus on the entire socio-technical system. This broader system could improve the decision-making process compared to directly following AI outputs, as humans may be able to consider whether the AI is less accurate in a particular case or have contradicting information that leads to a different decision (Buijsman & Veluwenkamp, 2022). On the other hand, operators could ignore the system, meaning that its performance is ultimately irrelevant for the outcomes of (and thus also our consent to) the decision making process, as happened in a pilot for detecting cardiac arrest in emergency calls, Zicari et al. (2021). This holds, similarly, for fairness as the interaction of the AI system with operators may change the ultimate distribution of decisions compared to those of just the AI in isolation. Dwork and Ilvento (2018) have nicely demonstrated that a composed system of components that individually are fair can still as a whole produce unfair outcomes, and that unfair components may be

combined to form a system that as a whole meets standards of fairness. Likewise, whether decision subjects have the option to effectively contest a decision will have an impact on how the decision making procedure affects them, and the tension between contestability and consistency of the procedure make it a relevant aspect to include. It is, in short, the system as a whole that is the right unit of consent and of transparency.

The second question is what information is needed in order to consent to a particular decision-making procedure that includes AI. Here, I consider two broad categories that are relevant: the impact that the system will have on the people that need to consent and the way in which the procedure itself takes their values and perspectives into account. For the first consideration, the reason is that we generally consider outcomes to be relevant in judging whether a procedure is acceptable or not. The procedure for a trial, for example, is often determined to be just based on how often it leads to the correct outcomes of convicting the guilty and releasing the innocent. This idea is found in other approaches to legitimacy and procedural justice such as in Rawls (2001) discussed further below, and generally fits the intuition that we determine consent in large part based on what the consequences are for us. The impact of the decision-making system is then the first thing that transparency should give information on.

Secondly, consent will likely depend on more than just the impact of the procedure in question. Empirical evidence suggests that people care about more than just outcomes when judging the legitimacy of a process. This has been the subject of philosophical work by Meyerson et al. (2021) and goes back to the empirical work started by Thibaut and Walker (1975). As already nicely summarized in Meyerson and Mackenzie (2018), this work shows first and foremost that people are more satisfied with outcomes that they receive when a process is perceived as fair regardless of whether the outcome is favourable to them. In other words, when experimenters ask people whether they are satisfied with an outcome they will look at whether they consider the procedure to be fair [which I will here interpret, in line with Meyerson and Mackenzie (2018), as amounting to the same thing as whether they see it as legitimate—under the assumption that they wouldn't consent to unfair procedures, or deem their outcomes of unfair procedures acceptable]. For this judgement of the fairness of a procedure they do not look at the accuracy of the process, but rather at whether the process appropriately considered their opinions and needs. "The evidence shows that people do not primarily associate just procedures with procedures that contribute to accurate outcomes. Nor do people evaluate the justice of procedures by reference to how well these procedures respect their capacity to reason. Instead, people evaluate procedures from an interpersonal and relational perspective, in terms of their capacity to enhance the quality of their interpersonal interactions with authorities, this being something that they value for its own sake" (Meyerson and Mackenzie 2018, p. 7).

In particular, this research highlights the people care about what is called"voice": the opportunity to be heard during the procedure or after outcomes have been finalized (Burke & Leben, 2008; Folger, 1977; Lind et al., 1990) as well as"benevolence", i.e. the willingness of decision makers to consider one's needs and the clear communication of this willingness (Blader & Tyler, 2003; Lind & Tyler, 1988). These general findings have been studied more specifically in the context of algorithms by Lee et al. (2019), who investigated people's perception of an algorithm based on different kinds of information and control over the decision making. 'Standards clarity', where the basic functioning of the algorithm (rules-based in this particular study) was explained to participants failed to improve their perception of the fairness of the algorithm. "Many understood how the algorithm worked and how fairness in division was operationalized based on this step-by-step description, and they later used the knowledge to interpret the input and output matrix table. At the same time, participants told us that the standards clarity alone did not make them trust the algorithm or see the results as being automatically fair." (Lee et al. 2019, p. 13) Outcome control, where participants got to experiment with changing outcomes, were able to discuss alternatives with the rest of the group and possibly deviate from the algorithm's recommendation, did significantly improve people's perceptions of fairness even though 80% of the groups made no changes to the allocation of goods. The interpretation here is that precisely aspects such as voice (being able to discuss) and benevolence (seeing that there were no better allocations available) contributed to this improvement. However, one may also interpret it as showing that participants realised that the outcomes were optimal, and legitimizing the procedure on that basis.

In another empirical study regarding the legitimacy of algorithms, Martin and Waldman (2022) found that when arbitrary or morally problematic grounds are behind an algorithmic decision, it is not viewed as legitimate regardless of whether the outcome is positive or negative. If the reasons for the decision were deemed in order, then algorithms leading to positive outcomes were judged as more legitimate than those leading to negative outcomes (decisions that negatively affected the subject or that she disagreed with). Now, this was only using single outcomes of the procedure, and thus doesn't rule out that we care about the reasons for the decision because the wrong reasons lead to more incorrect decisions overall. I therefore do not think that the empirical results specific to algorithms settle the question against a purely outcome-based account, but as it is compatible

with the broader set of evidence which arguably does conflict with outcome-based approaches I maintain this as a valid argument for algorithms as well. This is hardly surprising as these procedures do not seem fundamentally different from the (legal) processes that have been studied in the wider literature. Still, it would be good if further empirical studies would distinguish elements of voice and benevolence from the overall accuracy of the procedure specifically in the context of algorithmic decision making.

Wrapping up, there are three general requirements that stem from this discussion. First, transparency should consider the entire (socio-technical) decision-making process, not just the AI system in isolation. Second, transparency should help the relevant parties identify the impacts of the decision-making system. Third, transparency should highlight not just aspects of the outcome but should also inform on the purely procedural aspects of the decision-making process, in particular those of voice and benevolence. One may wonder, however, whether this analysis depends on the choice of consent-based views of legitimacy. Would a different approach to legitimacy lead to alternative requirements of transparency? While one doesn't get to all three requirements in most cases, I think that these three provide a happy medium between alternative accounts that pick up on the central points of all of them.

On one set of accounts the first two requirements for transparency are easy to derive. (Binmore, 2000) approaches legitimacy through the actual impact of a decision-making procedure. On strict versions of such a view, where only the actual impact on utility matters, it is easy to see that (a) the unit of analysis should be the entire socio-technical system (as this determines the actual impact on utility) and that the outcomes of this system matter. We need to know the impact of the decision-making procedure to determine whether it is legitimate. Related outcome-based views such as the Rawlsian view of (imperfect) procedural justice Rawls (2001) appealed to by Loi et al. (2021) in their discussion of legitimacy view procedures as just if they lead to the right outcomes often enough. Since the relevant outcomes are those of the entire decision-making procedure, we get the first point. The relevance of outcomes on these views was already mentioned above and is part of why I find it plausible that outcomes are normatively relevant and matter for (determining) the legitimacy of decision-making procedures. The only requirement posed above that does not directly flow from these types of views of legitimacy is the third one, highlighting the importance of purely procedural aspects of the decision-making procedure. I consider that the arguments made in favour of that importance nevertheless hold up, as they were not specifically consent-based but rather look at people's perceptions of legitimacy, coupled to a relational perspective.

The exact opposite holds for these relational and more pure procedural accounts of legitimacy, of which Peter (2009) is another example. On this conception, democratic legitimacy comes from a decision-making process satisfying conditions of epistemic and political fairness, regardless of outcomes. Here, too, the first requirement is quickly derived as the entire decision-making process will have to satisfy these requirements (and likewise, for Meyerson and Mackenzie (2018) we can only consider if voice and benevolence are sufficiently respected when looking at the process as a whole). Outcomes are less clearly important, but the procedural aspects of including people in the decision-making process in the right way are central to these accounts. Outcomes might not matter on these views, and so the second requirement for transparency is not one that is easily defended if one opts for purely procedural accounts.

Combining these different views on legitimacy we arrive, again, at the three requirements identified. A broad view on the socio-technical system is important on each of the discussed accounts. There is, of course, disagreement on whether outcomes of decision-making processes or purely procedural features are leading in determining if a process if legitimate or just. However, for requirements on transparency there is no need to resolve this debate. It highlights, rather, that both elements are relevant to the question at hand. To be sure that we can establish the legitimacy of a decision-making process we need to demonstrate both the impact that the process will have on the relevant people and the way in which the procedure itself is shaped, with the latter especially focused on how people are included in the procedure. With those ideas in mind, the next two sections consider current approaches to transparency in light of these requirements. They are, broadly speaking, split along the same lines as the views on legitimacy in terms of focusing either on the process itself or on the outcomes of the process.

One final point to consider here is what types of AI systems are covered by these requirements. Most of the views discussed here focus on political legitimacy, and the examples below (especially in Sect. "Current approaches: Process-based transparency") also focus on the public sector. This makes sense, as decision making by governments involves political power that is often lacking from decision making by companies (and we e.g. have alternatives if a bank denies a loan whereas there are no alternatives if social benefits are denied by the government). Therefore, accountability and the need to establish legitimacy is more important for the public sector (whether it develops AI systems on its own or through public-private partnerships, as long as the deployment is in the public sector) than it is for the private sector. Still, even if the normative obligations that attach to power are less applicable to companies there is good reason to think that transparency should be understood along similar lines. When companies have to show to regulators

that their decision-making processes are legitimate, the same considerations of outcome- and process-based factors will be relevant. Likewise, if companies want to convince their clients that they are using AI in an acceptable manner the way in which to show this will be the same. The difference is, rather, one of degree: companies might have a lower bar to meet before the use of an AI system is legitimate (because it's not using public funds and because it typically has less power over citizens). The relevant criteria, however, do not change. With that in mind, the next question is how these criteria relate to current approaches to transparency.

Current approaches: process-based transparency

One type of information that can be given regarding AI systems is the way in which the development process took place. What data was gathered, how was this processed and what type of model was trained based on that data? What checks were done for risks such as algorithmic bias and how was the AI system implemented in the wider decision making procedure? Answers to these questions need not mention the actual performance of the algorithm, as is well illustrated by some of the entries in the algorithm registers that are being experimented with in the Netherlands. When considering for example the (April 2024 version) of the entry for a system that estimates the risk for illegal holiday rental (https://algoritmeregister.amsterdam.nl/en/illegal-holiday-rental-housing-risk/) we find the following types of answers after a general overview of the reason the systems was tested. On the accuracy of the system there is the following heading:

> **Performance**
> The advantage of a"random forest regression"is that it is a fairly complex algorithm that can approximate reality quite well. However, there is a risk of overfitting. A"tree"with many layers squeezes the data to provide specific answers. It has been researched how many layers the model needs to remain generic and therefore, not to overfit. In addition, continuous data points are categorized (grouped), so that the model has a clear number of options instead of the infinite number of continuous values. This makes the model better suited to reach a conclusion.

Likewise, "To make sure employees understand the consideration that the algorithm is making, the"SHAP"method is used (SHapley Additive exPlanations: https://github.com/slundberg/shap). SHAP calculates, which features in the data have resulted in high or low suspicion of illegal housing. This ensures that an employee can always understand what the algorithm based its risk assessment on, so they can make a well-considered decision."As well as under non-discrimination the information that "a group can still be disadvantaged by the algorithm, even if the group is not explicitly known to the algorithm. We have therefore chosen to conduct further research into this form of algorithmic bias during the pilot. For this we use the"AI Fairness 360 toolkit"(https://aif360.mybluemix.net)."Finally, under the heading human oversight, "A work instruction has been drawn to prevent employees from having excessive confidence in the algorithm. In addition, the employees undergo training to recognize the opportunities and risks of using algorithms."(all quotes are from https://algoritmeregister.amsterdam.nl/en/illegal-holiday-rental-housing-risk/)

All of these pieces of information show that the municipality has actively considered risks that come with the development of AI systems and is taking steps to mitigate these risks as much as possible. This is clearly relevant information, and much more than what is typically disclosed about the use of algorithms by organisations. What is interesting, however, is that there is no mention of the effectiveness of any of these measures. The developers aimed to prevent overfitting, but the algorithm register doesn't give any information on the accuracy of the algorithm on new data. Likewise, steps are taken to improve the explainability of the algorithm, but how well this works is not mentioned. The same goes for non-discrimination and human oversight.

This is a problem, as it gives only limited insights into the functioning of the algorithm due to the unpredictability of these algorithms. The text on performance can apply just as well to an algorithm that has 60% accuracy as to one that has 99% accuracy. Likewise, the use of SHAP does not guarantee that the system is explainable (in the sense that the reasons for individual outputs are understood by users), and in fact the academic literature has so far found only marginal improvements in objective measures such as the ability to predict outputs or to make correct decisions (Chromik et al., 2021; Wang & Yin, 2021) based on SHAP. The same points can be made regarding non-discrimination, where the different fairness measures shown in the IBM toolkit are known to conflict (Kleinberg et al., 2016) and choices between these different measures (e.g. balancing true positive rates v.s. balancing false postive rates for specific groups) have a clear impact on the outcomes. A particularly clear example of this is the simulation study run by Liu et al. (2018) where we see that financial inequality can increase over time between two groups even when an algorithm predicting credit risk is fair in the sense that both groups get loans equally often. The reason being here that the equal distribution of loans can lead to the poorer group getting proportionally more loans that they cannot afford, thus defaulting more often. As defaults are costly, financial inequality can grow over time despite the algorithm being fair according to the widely used demographic parity

measure. In short, the choice of fairness metric matters a lot, and without any information on this choice it is hard to say whether the algorithm meets non-discrimination standards.

How does this stack up to the transparency requirements based on the goal of establishing legitimacy? First, this example has taken a more socio-technical approach to transparency as human oversight is explicitly included in the reported features of the system. On the other two requirements, it is however less successful. This kind of process-based approach to transparency, which informs you which steps are taken but not what the outcome is of these steps, fails to inform us on both the impact of the decision-making system and on how successfully citizens voice is heard and considered. The behaviour of algorithms, and machine-learning algorithms in particular, is difficult to predict and good intentions are insufficient to ensure a working algorithm [c.f. the failure of hundreds of machine-learning algorithms developed to diagnose covid, Heaven (2021)]. As discussed above, steps to improve explainability and fairness also don't guarantee outcomes that we find acceptable, if we don't know what their effectiveness was or which trade-offs were made. Ultimately, one has to trust that the developers saw outcomes that they deemed acceptable, rather than that one is able to determine for oneself whether this kind of algorithm is justified for this kind of goal. As such, important information is missing to establish legitimacy. Can current outcome-based approaches do better?

Current approaches: outcome-based transparency

The current standard for algorithm transparency in the computer science community is arguably set by Mitchell et al. (2019) in the form of model cards. These consist of long lists of information about an algorithm, organised into Model Details, Intended Use, Factors, Metrics, Evaluation Data, Training Data, Quantitative Analysis, Ethical Considerations and Caveats & Recommendations. Each of these sections has a number of sub-questions to be answered by the organizations that have developed the model, and they ultimately are supposed to give information on the performance of the model on the relevant factors (i.e. both general performance but also differences in performance between relevant sub-groups such as between men and women). The earlier sections identify what the relevant sub-groups are, as well as what was measured for these different sub-groups and on what data. In other words, model cards are focused not on informing us about the steps that were taken in the design of the algorithm, but rather focus on listing a number of (primarily quantitative) properties of the finished algorithm. These properties can then be judged, so is the thought, based on the intended use of the model. Do we consider that it has the right features to fulfill that intended use? Ideally that question can be answered based on the quantitative analysis, ethical considerations and caveats section. Whether this happens in practice depends very much on who fills in these model cards (Boyd, 2021; Heger et al., 2022), but that is an issue that I will set aside.

For the more conceptual issues under discussion here it is interesting that the model cards approach aligns quite closely with the Design Publicity framework laid out by Loi et al. (2021). On this account, there are four types of transparency that should be offered with respect to an algorithm: value transparency, translation transparency, performance transparency and consistency transparency. Briefly put, these consist of the goal of the algorithm (the values), how this goal is formulated in mathematical terms (translation), what the scores are on a set of metrics relevant to the mathematical formulation of the goal (performance) and how consistently these scores are maintained over re-training etc (consistency). In their own words:

> The goals or values that guide the design of algorithmic models should therefore be included in an explanation of such models. *Value transparency* is the result of an explanation that makes the standards, norms, and goal that were implemented in the system accessible. These normative elements should also correspond to the *reasons for which* it was deployed. (Loi et al. 2021, p. 257)

For what follows it is important to note that the values and goals here are understood in terms of the problem that the algorithm is supposed to solve and as such matches fairly well with the Intended Use section of the model cards approach. For machine learning algorithms this is typically captured in the loss function that is optimized for during the training phase and in the way data has been labelled, but for other types of algorithms it will be less explicitly present. In any case, the fact that they are thinking about values that are explicitly translated into the code of the algorithm becomes clearer in the discussion of translation transparency:

> The goal of an algorithmic system needs to be translated into something that is measured: a set of rules with which the algorithm elaborates inputs and produce outputs. ... There is no straightforward and unique way to translate a goal into a mathematic construct. ... For this reason, making such translation a publicly verifiable criteria provides the public and scientific community with the information to assess how a given goal is operationalized in machine-language. (Loi et al. 2021, p. 257)

As an example, they point to the concept of customer churn which could be defined in different ways (meaning that the

labelling of data points as showing customer churn can happen in different ways), for example as a customer that hasn't spent anything for a year with the company or as a customer that hasn't made use of any services from the company for a year (thus excluding those who used free promotions during the year from the definition of churn). This idea is similar to that argued for by Casacuberta et al. (2022), who use the term 'justificatory explanations' for information about the way key concepts used by the algorithm are understood/were defined. In the model cards this is ideally (as in one of the examples in the paper) described under the Data section, where for an algorithm classifying toxicity in text it is listed that ""Toxic"is defined as"a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."" (Mitchell et al. (2019), p. 227). In short, the way in which the goal, and particularly key concepts in that goal, have been understood by developers should be communicated clearly. That will then help to place the quantitative evaluation of the algorithm, which is done using the mathematical version of these concepts.

> Performance transparency consists in indicating the logic with which the algorithm has been tested in order to verify how much it departed from achieving the goal and in indicating the results of such logic, starting with the choice of performance measures used in both training phase and during the assessment of the model on test data (Loi et al. 2021, p. 258)

The examples given in accompaniment refer to the accuracy and the fairness of an algorithm, both of which can be measured based on the outcomes of the algorithm run on test data. As such, there is again a strong overlap with the model cards approach, which lists a quantitative analysis over specified test data and with fairness metrics worked out for the sub-groups identified in the Factors section. In terms of content, then, there isn't much of a difference between the two approaches. Model cards have an extra field for additional ethics concerns, recommendations and caveats, whereas Design Publicity additionally requires information about the consistency of the algorithm. "Consistency transparency is showing proof that consistency is achieved, i.e. that the algorithm always generates predictions by the same rules even when we cannot observe those rules in operation."(Loi et al. 2021, p. 259)

Not only do we have an interesting overlap between computer science and philosophical approaches to transparency here, there is also a good philosophical basis for the account developed by Loi et al. (2021). They use Rawls' notion of (usually imperfect, given that algorithms are rarely perfectly accurate) procedural justice, where a procedure is just if it leads to the right outcome often enough. Individual decisions are then justified if they are the result of a just procedure that has been consistently applied (hence the addition of consistency transparency). It is also why the information is focused on whether the goal is acceptable and whether the algorithm fulfills that goal well enough. The ideas are not as explicit in the model cards, but the same idea that of wanting to know what the algorithm is supposed to do and how well it does it is behind that framework.

Generally speaking I am sympathetic to this approach, and the alternative I will suggest in Sect. "Current approaches: Outcome-based transparency" is in essence an extension of these two outcome-based approaches. However, when comparing the approach to the requirements on transparency from Sect. "What kind of transparency do we need?" there are two important parts that I think are missed by both Mitchell et al. (2019) and Loi et al. (2021). First, they do not include broader socio-technical elements of the decision-making procedure that are important to the outcomes but fail to be captured by performance measures of the algorithm alone. Second, the more purely procedural aspects of the decision-making procedure are left out of this approach to transparency.

To expand a bit on both points in relation to the specific suggestions by both approaches, there is first of all no mention of human oversight or the contestability of the decision-making procedure. In fact, the entire interaction with human decision-makers is left out from the frameworks. Furthermore, there is a risk that the focus on outcomes gets reduced to a focus on the accuracy and fairness of the algorithm [which, not to forget, also gives a limited view on fairness as the fairness of the composed decision-making procedure can diverge significantly from the fairness of the AI system itself (Dwork & Ilvento, 2018)]. This is certainly what one sees in the quantitative analysis samples from Mitchell et al. (2019), which list different accuracy scores both in aggregate and for the relevant sub-groups and leave it at that. The Rawlsian approach of Loi et al. (2021) tends towards the same direction, as getting to the right outcome often enough could be interpreted as ensuring that the algorithm is sufficiently accurate, and not disproportionately inaccurate for certain groups. While the authors do mention other considerations, such as privacy [which could be measured using e.g. k-anonimity and l-diversity, Machanavajjhala et al. (2007)] there is no clear reason why this would be needed from the Rawlsian rationale they adhere to as applied strictly to the algorithm's outcomes. It is also not sommething that pops up in the examples they give, and so while there is no principled reason why these other values are excluded by the design publicity framework there are also no indications in the theoretical work that they should be included.

Hence my worry that values other than accuracy and fairness are overlooked in the current outcome-based

approaches to transparency and that as a result one misses the outcomes that truly matter, namely those of the entire socio-technical system. That being said, a broader view of the decision making procedure and the effects it has on society would show that outcomes which we deem unacceptable occur more frequently if people's privacy is not preserved [e.g. identity theft, mistreatment based on the available personal information and threats to autonomous decision making; DeCew (2018)]. As such, one could argue that if we look at the general consequences of using the algorithm then these considerations neatly come back into the purview of design publicity and the Rawlsian rationale behind it. I won't dispute this, but do think that it's important to keep in mind that for that a broader view of what transparency is about [not just the algorithm itself, but the socio-technical system in which it is implemented and of which the data is a part) is needed. That broader view is missing in the design publicity framework as it is presented (model cards intentionally do not cover privacy, as that is covered in the datasheets framework, Gebru et al. (2021)].

Second, accounts of transparency that are fully geared towards showing how often an algorithm arrives at the right outcomes miss the purely procedural aspects of decision-making that are also relevant for establishing legitimacy. To go back to the relational approach of Meyerson and Mackenzie (2018), the involvement of people in the process is central to procedural justice. "A commitment to procedural justice entails a commitment to processes that support self-respect by conveying respect for individual citizens as moral equals... This requires processes that demonstrate officials' impartiality, willingness to listen, commitment to justice, and concern for the welfare of those who appear before them."(Meyerson and Mackenzie, 2018, p. 9) Neither of the two outcome-based approaches to transparency discussed here pay attention to the involvement and relation to people in the decision-making process. As such, they miss an important element to the legitimacy and acceptability of AI systems. In the next section I propose a value-based approach to transparency that manages to pay attention to both the impact of the socio-technical system and the procedural aspects of the decision-making procedure.

Value-based transparency

The starting point for my account of transparency is the general view of ethics of technology that comes with Design for Values (Van den Hoven et al., 2015) and Value-Sensitive Design (Friedman et al., 2002, 2013). On these approaches to technology the idea is that technologies, including algorithms, are not neutral. On the contrary, values (often those of the designers) become embedded into technologies during their design (van de Poel, 2020), resulting in e.g. safety by design (van de Poel & Robaey, 2017) and privacy by design (Gürses et al., 2011). The way in which this happens, ideally, is that after determining the relevant values for a technology under development these are further specified into norms. These more detailed specifications show what aspects of these values filter into the technology (e.g. privacy can be both about preventing harms such as identity theft and about preserving autonomy). Finally, these norms are translated into design requirements that, if satisfied, would realize the value in the technology (though this should be verified in case the design requirement didn't fully capture everything we wanted). Along the way there is the tricky matter of conflicting values, such as fairness and accuracy, where trade-offs or innovation is needed to either choose or find a way to realize both values despite the initial conflict (Van de Poel, 2009).

This idea is now widespread, and talk of values can be found e.g. in the EU High-Level Expert Group's guidelines for AI Ethics where seven overarching values are listed that should guide the development and use of artificial intelligence. My claim here is that we can use the same idea for determining what transparency about algorithms should look like. If we take the different values that we hold for the socio-technical system of which the algorithm is a part (to not forget about contestability, human oversight and the effect of human intervention on values such as fairness) then what transparency should do is inform what these values are and how successfully they have been realized in the system. To do so, it will be important to show how the values have been translated into design requirements/what measures are used to determine if the system is fair, reliable, safe, etc. I thus propose that transparency for algorithms be organised in the following way:

1. An overview of the different values identified as relevant for the socio-technical system
2. A conceptual specification of these values (e.g. privacy, health, fairness is understood as x)
3. A quantification of these specifications values into verifiable design requirements
4. The performance on the measures formulated under (3), including how consistently this performance is maintained over time
5. An elucidation of any value conflicts, and the choices made in light of these conflicts

For example, a fraud detection system would have accuracy (i.e. preventing fraud) as a central value, but would also be designed to consider fairness, contestability, and more. Some of these values will be required by third parties, such as the European Union, who help to democratically set the standards for when a decision-making procedure is legitimate. However, not all the relevant values will

have been identified by such third parties and while legal requirements are a clear minimum for a procedure to be legitimate there are often cases that are legal but still (normatively) unacceptable. I therefore consider values such as those listed by the High-Level Expert Group to provide a good starting point, but one that in practice organizations will have to expand with values relevant to the specific context. That gives the initial overview for step 1. Step 2 then explains how these values have been understood. Fairness could for example be defined as an equal distribution of inaccurate accusations of fraud, whereas contestability could be the availability of an accessible and effective mechanism to correct inaccurate accusations in a way that decision subjects have minimal negative consequences of the mistake. Then, measures can be formulated for these different values, e.g. equal false positive rates computed over both the algorithm itself and the final decisions by the human operator as well as accessibility of the contestation mechanism, number of mistakes redressed and damage to the accused. Point four would then detail the scores on these measures and point five could discuss the conflict between identifying fraud versus avoiding incorrect accusations of fraud, among others. This can help justify why the scores will not be perfect on all performance measures, as trade-offs have to be made between different values.

In this sense, the proposal is outcome-based as the framework drives towards the performance on measures associated with the different values. The difference is that these measures can be about more than the output of the algorithm, which makes it diverge from the focus on accuracy and fairness (which in the examples in Sect. "Current approaches: Process-based transparency" tends to boil down to differences in accuracy between groups) and allows room for properties of the procedure itself. This neatly solves the two issues that I pressed in Sect. "Current approaches: Outcome-based transparency", as well as the worries about process-based transparency from Sect. "Current approaches: Process-based transparency". To briefly address the latter, the worry there was that if transparency focuses on the steps taken during development, then we don't know enough about the actual decision-making procedure to judge whether it is legitimate or not. While my framework ideally does reflect the choices made during the development (namely that values were identified, specified and designed for) it also informs on the end result thanks to item number 4. Properties of the algorithm are included, and thus people would know how effective the steps taken have been. At the same time, by explicitly listing which values have been considered during the design and implementation of the algorithm it should be possible for e.g. investigative journalists or ethics auditors to spot when important values have not been considered. Of course, the algorithm need not be problematic (one could comply even if not by design), but it would be a reason to question the legitimacy of using the algorithm until steps have been taken to verify the system on these points.

There is an equally straightforward resolution of the issues from Sect. "Current approaches: Outcome-based transparency". First, there is the problem that broader socio-technical aspects were not included in current approaches. As values are broadly defined here, they will include human oversight, contestability, etc. As an example, this could include measures of the number of mistaken AI outputs that are corrected by human operators (and vice versa, the number of correct AI outputs that the operators end up disagreeing with) as well as the number of decisions that are contested (successfully). Performance for these, and other measures, should also be measured over the outcomes of the socio-technical system as a whole, though it is both relevant and easier to compute some performance measures of the algorithm in isolation.

Second, we have to deal with the empirical findings that aspects of the procedure such as voice and benevolence are important to (perceptions of) legitimacy. Voice is naturally treated as an additional value, closely linked to contestability, that highlights the importance for people to be involved in the decision making process—even when algorithms are a part of how decisions are made. Benevolence, on the other hand, should follow from the overall approach. By showing participants what values have been considered and how, values that should align (to a good extent) with what is important to them, it should already become clear that their interests are considered in the decision making process. Of course, one could include benevolence as an additional value as well—the general framework may not cover more specific things such as how human operators treat unintentional mistakes—but overall the attitude to decision subjects is likely to become clear from the different values considered and how value conflicts are handled. In short, we can take these empirical findings as telling us something about what people value in (high stakes) decision making procedures, and as such incorporate it into the aspects that we should be transparent about (and should design for).

As a result, a value-based approach to transparency manages to combine procedural and outcome considerations for legitimacy. It is, thus, a way to meet all three the requirements on offering transparency about the use of AI systems that were listed in Sect. "What kind of transparency do we need?". It is based on the broader socio-technical system, shows the impacts of the decision-making procedure on decision subjects and highlights how procedural/relations features such as voice and benevolence are taken into account. This combination is possible because the impact of the system is broadly construed, to mean impact on what we find valuable. As this includes both procedural and outcome-based features this approach naturally combines the two.

It also closely mirrors the development process, at least in ideal cases. AI systems should be developed in accordance with the values of the different stakeholders that are impacted by them. To ensure that this happens we need a systematic approach to the inclusion of these values. In that sense, the transparency requirement here asks for no more than that organizations that develop and implement AI show how they have incorporated (ethical) values during the development of the (socio-)technical system and that they report how successful this was. At the same time, if this kind of reporting proves difficult for an organization then that can be a signal in itself that important values are not explicitly considered and safeguarded during development. What one has to be transparent about on my account, after all, are precisely the consequential choices that have been made in the design, and (in addition to the Sect. "What kind of transparency do we need?" approach) how those choices have turned out in practice. This helps signal if those choices are made consciously, as opposed to driven purely by the requirement of having a system that is as accurate (or, e.g. profitable) as possible. At the same time it can link to a Rawlsian procedural justification for a socio-technical system, by showing that it leads to acceptable (taken here not as accurate, but as in line with our wider set of values) outcomes often enough. Thus, a values-based approach to transparency that maintains a focus on the performance along measures stemming from those values is the best way forward.

Conclusion

What information should be provided to the general public about algorithms? I've argued that what matters is the values that have been embedded in the socio-technical system of which the algorithm is a part. This takes a more outcome-based stance than the transparency frameworks discussed in Sect. "Current approaches: Process-based transparency", where the focus is on the steps that have been taken by developers during the design of the algorithm as well as the broader socio-technical system. At the same time, it takes a broader view than the current outcome-based transparency frameworks discussed in Sect. "Current approaches: Outcome-based transparency", which risk reducing the provided information about a system to the accuracy and statistical fairness of the algorithm. As such, a value-based framework satisfies the three requirements on transparency that were outlined in Sect. "What kind of transparency do we need?": (1) consideration of the broader socio-technical system, (2) showing the impacts of the socio-technical system on stakeholders, (3) attention to the purely procedural aspects of decision-making procedures.

By being clear about the different values that (should be) realized in the system it is possible to show the (ethical) considerations that went into the design and decision to implement the system. By detailing how these values were translated into design requirements (via the intermediate step of conceptual specifications) it is, moreover, possible to show how successful these intentions have been. I believe that by doing so we can provide the public with all the information that is needed to determine whether a decision making procedure is legitimate.

Data availibility Data sharing not applicable to this article as no datasets were generated or analysed during the current study**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Amnesty International. (2021). Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. Technical Report EUR 35/4686/2021.

Binmore, K. (2000). *A Utilitarian Theory of Legitimacy, in Economics, Values, and Organization*. Ben-Ner, Avnerand Louis G. Putterman (eds.), Cambridge: Cambridge University Press, pp. 101–132.

Blader, S. L., & Tyler, T. R. (2003). A four-component model of procedural justice: Defining the meaning of a "fair" process. *Personality and Social Psychology bulletin, 29*(6), 747–758.

Boyd, K. L. (2021). Datasheets for datasets help ml engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–27.

Buijsman, S. (2022). Defining explanation and explanatory depth in xai. *Minds and Machines, 32*(3), 563–584.

Buijsman, S., & Veluwenkamp, H. (2022). Spotting when algorithms are wrong. *Minds and Machines, 33*, 541–562.

Burke, K., & Leben, S. (2008). Procedural fairness: A key ingredient in public satisfaction. *Court Review, 44*, 4–25.

Casacuberta, D., Guersenzvaig, A., & Moyano-Fernández, C. (2022). Justificatory explanations in machine learning: For increased transparency through documenting how key concepts drive and underpin design and engineering decisions. *AI & society, 39*, 279–293.

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th international conference on intelligent user interfaces*, (pp. 307–317).

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.

DeCew, J. (2018). Privacy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2018 edition.

Dwork, C., & Ilvento, C. (2018). Fairness under composition. arXiv preprint arXiv:1806.06122.

Estlund, D. (2009). Democratic authority. *Democratic authority*. Princeton University Press.

Folger, R. (1977). Distributive and procedural justice: Combined impact of voice and improvement on experienced inequity. *Journal of Personality and Social Psychology, 35*(2), 108.

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington technical report,* 2, 12.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory* (pp. 55–95). Springer.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

Greene, A. R. (2016). Consent and political legitimacy. *Oxford Studies in Political Philosophy, 2*, 71–97.

Gürses, S., Troncoso, C., & Diaz, C. (2011). Engineering privacy by design. *Computers, Privacy & Data Protection, 14*(3), 25.

Heaven, W.D. (2021). Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review*. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covidhospital-diagnosis-pandemic/. Accessed 30 July, 2021.

Heger, A. K., Marquis, L. B., Vorvoreanu, M., Wallach, H., & Wortman Vaughan, J. (2022). Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–29.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint http://arxiv.org/abs/1609.05807.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction, CSCW*(2), 1–26.

Lind, E. A., Kanfer, R., & Earley, P. C. (1990). Voice, control, and procedural justice: Instrumental and noninstrumental concerns in fairness judgments. *Journal of Personality and Social psychology, 59*(5), 952.

Lind, E. A., & Tyler, T. R. (1988). *The social psychology of procedural justice*. Springer.

Liu, L.T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning*, (pp. 3150–3158). PMLR.

Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and Information Technology, 23*(3), 253–263.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD), 1*(1), 3-es.

Martin, K., & Waldman, A. (2022). Are algorithmic decisions legitimate? The effect of process and outcomes on perceptions of legitimacy of AI decisions. *Journal of Business Ethics, 183*, 653–670.

Meyerson, D., & Mackenzie, C. (2018). Procedural justice and the law. *Philosophy Compass, 13*(12), e12548.

Meyerson, D., Mackenzie, C., & MacDermott, T. (2021). *Procedural justice and relational theory: Empirical, philosophical, and legal perspectives*. Taylor & Francis.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, (pp. 220–229).

Peter, F. (2009). *Democratic legitimacy*. Routledge.

Peter, F. (2017). Political Legitimacy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2017 edition.

Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

Simmons, A. J. (2001). *Justification and legitimacy: Essays on rights and obligations*. Cambridge University Press.

Thibaut, J. W., & Walker, L. (1975). *Procedural justice: A psychological analysis*. L. Erlbaum Associates.

Van de Poel, I. (2009). Values in engineering design. *Philosophy of technology and engineering sciences* (pp. 973–1006). Elsevier.

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines, 30*(3), 385–409.

van de Poel, I., & Robaey, Z. (2017). Safe-by-design: From safety to responsibility. *Nanoethics, 11*(3), 297–306.

Van den Hoven, J., Vermaas, P. E., & Van de Poel, I. (2015). *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*. Springer.

Varošanec, I. (2022). On the path to the future: Mapping the notion of transparency in the EU regulatory framework for AI. *International Review of Law, Computers & Technology,*. https://doi.org/10.1080/13600869.2022.2060471

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *26th international conference on intelligent user interfaces*, (pp. 318–328).

Zicari, R. V., Brusseau, J., Blomberg, S. N., Christensen, H. C., Coffee, M., Ganapini, M. B., Gerke, S., Gilbert, T. K., Hickman, E., Hildt, E., et al. (2021). On assessing trustworthy AI in healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Frontiers in Human Dynamics*. https://doi.org/10.3389/fhumd.2021.673104

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.