AI & SOCIETY (2022) 37:375–381
https://doi.org/10.1007/s00146-021-01200-5

OPEN FORUM

Algorithmic and human decision making: for a double standard of transparency

Mario Günther[1] · Atoosa Kasirzadeh[1,2]

Received: 16 May 2020 / Accepted: 22 March 2021 / Published online: 6 April 2021

Abstract

Should decision-making algorithms be held to higher standards of transparency than human beings? The way we answer this question directly impacts what we demand from explainable algorithms, how we govern them via regulatory proposals, and how explainable algorithms may help resolve the social problems associated with decision making supported by artificial intelligence. Some argue that algorithms and humans should be held to the same standards of transparency and that a double standard of transparency is hardly justified. We give two arguments to the contrary and specify two kinds of situations for which higher standards of transparency are required from algorithmic decisions as compared to humans. Our arguments have direct implications on the demands from explainable algorithms in decision-making contexts such as automated transportation.

Keywords Algorithmic decision making · Transparency · Explainable AI

1 Introduction

Nowadays algorithms are used to make impactful decisions. Algorithms recommend whether or not access to credit is granted (Davis et al. 1992), whether a job application is rejected (Gonzalez et al. 2019), or who receives which medical treatment (Obermeyer et al. 2019). Algorithms are even used for assisting judges to pronounce judicial sentences. Quite naturally, and especially when the stakes are high, people would like to understand how the decisions informed by algorithms (as in the case of judicial sentencing) or directly made by algorithms (as in the case of loan application in some banks) come about. There is thus a growing demand for transparency to trace back the reasons for the algorithmically informed decisions.

The decision-making algorithms are designed by Artificial Intelligence (AI) researchers. The best of these algorithms are often based on Machine Learning (ML) techniques. ML algorithms do not encode a set of specific rules designed by the programmer to solve a class of problems. Rather such algorithms learn hidden patterns and structures from observed data about how to perform the task at hand, and then show some success in making accurate predictions of unobserved data in some domains. Despite this predictive accuracy, many ML algorithms are extremely complex and consequently opaque—even for their designers. This means that it is difficult for humans to understand the underlying reasons for the different algorithmic outcomes.

How should the transparency desideratum for algorithmic decision making be understood? One promising proposal is to compare the standards of transparency between human beings and algorithms (Zerilli et al. 2019; de Fine Licht and de Fine Licht 2020; Walmsley 2020). The way this dispute is settled would have resounding impacts on what to expect and how to design explainable algorithms which would make algorithms more transparent. There is a spectrum of positions—implicitly or explicitly argued for and narrow or broad in the scope of applicability—against a double standard of transparency. Some authors, such as Zerilli et al. (2019), make a strong claim that a double standard of transparency is hardly justified. Others, such as de Fine

Mario Günther and Atoosa Kasirzadeh contributed equally to this paper.

✉ Atoosa Kasirzadeh
Atoosa.Kasirzadeh@anu.edu.au

Mario Günther
Mario.Guenther@anu.edu.au

[1] Australian National University, Canberra, Australia

[2] University of Toronto, Toronto, Canada

Licht and de Fine Licht (2020), suggest that a double standard of transparency is not always required. Still others, such as Feller et al. (2016) and Corbett-Davies et al. (2017), seem to suggest that human and algorithmic decision makers are on a par with respect to fairness (and perhaps transparency) desiderata.

In this paper, we put forth two arguments for how and when a double standard of transparency is justified. Our arguments have direct implications for what we should demand from explainable algorithms in decision-making contexts such as automated transportation. To contextualize our position, we respond to Zerilli et al's. recent and influential position against the double standard of transparency. The main reason for this exposition is that, to the best of our knowledge, Zerilli et al's. paper offers the strongest arguments against a double standard of transparency. They observe that we demand from decision-making algorithms a higher degree of transparency than from human beings. They explain their observation: ML systems are held to higher standards of transparency *because* the transparency of human decision makers is overestimated. "The crucial point" in their argumentation "is that the standards of transparency […] can and […] *should* be applied consistently across the board, regardless of whether we are dealing with machines or humans" (p. 678). They conclude that the observed double standard regarding transparency is hardly justified. We give two arguments for how and when a double standard of transparency is justified. But, first, we review Zerilli et al.'s position in more detail.

2 A double standard for transparency is hardly justified

As Zerilli et al. (2019, p. 661) acknowledge, transparency is in itself an important desideratum: consequential decisions should be made as transparent as feasible, at least if there is no or only a relatively low cost in doing so. Hence, we have a default reason to make a decision system, be it human or algorithmic, as transparent as feasible without incurring too high of a cost. A fortiori, if it is feasible to make algorithms more transparent than humans, we have a default reason to do so.

Zerilli et al. (2019, p. 668) observe that we already hold algorithms to a higher standard of transparency than human beings. They swiftly explain why we impose this double standard. The reason be that human decision making appears more transparent than algorithmic decision making. They argue that this appearance is deceptive: human and algorithmic decision making are on a par as regards transparency. If their argument is correct, the putative reason for the double standard is undermined. And so they conclude that setting a double standard is hardly justified.[1]

How do Zerilli et al. establish that human decision making is not more transparent than algorithmic decision making? They argue that humans and algorithms alike suffer from the same kinds of bias. Thus bias alone cannot be the decisive factor for imposing higher standards of transparency for algorithms. So what makes human decision making appear to be more transparent? Zerilli et al. argue that humans can readily give practical reasons to justify their decisions. Yet the given practical reasons do not render transparent the underlying cognitive processes which led to the decision. While unarticulated intuitions and hunches play their role in the human process of arriving at a decision, we do not require that the corresponding opaque brain processes are made transparent. Zerilli et al. (2019, pp. 663–7) claim that ML algorithms, such as Deep Neural Networks (DNNs), are likewise opaque. But for DNNs we demand that their design and inner workings are made transparent. So, while we require higher standards of transparency from ML algorithms, we are fine with people just citing their practical reasons.

Zerilli et al. believe that practical reasons are sufficient for humans to justify their decisions. And since human decision making appears to be the gold standard for transparency, we would not need to go beyond the realm of practical reason to justify decisions. From this, they conclude that decision systems should—in general—not be held to higher standards than citing practical reasons. But what are practical reasons?

Zerilli et al. spell out what practical reasons are in terms of Dennett's (1987) *intentional stance*. There, Dennett maintains that the behavior of a system can be explained from three distinct stances. From the *physical* stance, we explain the behavior in terms of the fundamental sciences such as physics. From the *design* stance, we explain the behavior in terms of design principles. A computer program, for instance, can be explained on the design level without any need to explain how it works on the physical level. From the *intentional* stance, we explain the behavior purely in 'mentalistic' terms. The intentional stance is thus a level of analysis which abstracts away from the other two levels.

When we adopt the intentional stance, we use 'mentalistic' terms such as 'believe', 'desire', 'intend', and 'decide' to understand, explain, and predict the behavior of some

[1] Zerilli et al. do not consider different reasons that might justify a double standard. One such reason might be that, unlike machines, human beings have a right to privacy and so are protected from intrusive forms of transparency. It may turn out that AI systems are perhaps not demanded to be transparent because the transparency of human decision making is overestimated – but because humans enjoy rights machines do not. Here, however, we will not develop this possibility any further.

system, be it a human, or a chess computer for example. From the intentional stance, the behavior of a system is treated as if the system were a rational agent. That is, the entity is treated as if its 'decisions' are guided by its 'intentions' and 'reasons', perhaps in terms of its 'beliefs' and 'desires'. The scare-quotes indicate that some connotations of the words may be set aside. Yet they also point to a central role the terms play in practical reasoning: we attribute 'mentalistic' states to certain systems based on which their behaviour can be predicted.

Dennett (1987, p. 15) describes the intentional stance as a success notion. A system is amenable to an intentional stance explanation only if "its behavior is reliably and voluminously predictable" from the intentional stance. Moreover, we can ascribe a 'belief' that proposition *p* to an intentional system only if the most predictive interpretation of the system's behavior postulates that the system 'believes' that *p* (ibid., p. 29). Hence, an intentional system has a 'belief' that *p* only if this 'belief' helps in successfully predicting the system's future behavior. Dennett is clear that an explanation from the intentional stance must give us predictive power we did not antecedently have by adopting the other two stances. If the intentional stance does not provide additional predictive power, there is no intentional stance explanation in the first place (ibid., p. 23).

According to Zerilli et al., the only explanations we should require from human and algorithmic decision systems alike are intentional stance explanations. And since no more is needed, intentional stance explanations are to be preferred over those on the design or physical level. One need not know the technical details, for example how many nodes and how many hidden layers a DNN has, to explain why the ML algorithm decided so rather than so. We should ideally be able to trace back the 'reasons' for the outcomes. For then we could determine when to trust the AI and when the AI should be distrusted.

To sum up Zerilli et al.'s claim: intentional stance explanations are just right for justifying decisions, including algorithmic 'decisions'. An explanation for an algorithm's decision should thus be made from the intentional stance which abstracts away from the algorithm's design.

3 Counter-arguments

3.1 Argument from determination

As we have just seen, Zerilli et al. (2019) claim that algorithmic decisions should be explained from the intentional stance, and—since we do not expect more from human beings—only from there. This implies that intentional stance explanations should be preferred over design explanations. In their words on p. 661:

> We [...] argue that since the justification of action for human purposes takes the form of intentional stance explanation, the justification of algorithmic decisions should take the same form. In practice, this means that the sorts of explanations for algorithmic decisions that are analogous to intentional stance explanations should be preferred over ones that aim at the architectural innards of a decision tool.

Here, we argue that there are cases of algorithmic decisions where design explanations should be provided and indeed perhaps preferred over intentional stance explanations. We will do so by means of an example where the design explanation of an algorithmic decision should not only be taken into account but also preferred.

In 2017, Boeing 737 Max 8 s aircraft crashed during Lion Air Flight 610, killing 189 people (Johnston and Harris 2019). This aircraft has an algorithmically controlled stability system for adjusting the angle of the airplane. Faults with the design of this algorithmic system are reported to be among the main reasons for why this crash happened. In particular, the design was heavily relying on one sensor. This sensor malfunctioned and so sent an inaccurate signal about the airplane's angle. As a result, the algorithmic system pushed down the airplane's nose, and the airplane crashed.

From the intentional stance, the algorithmic system falsely 'believed' that the airplane was going up (due to the inaccurate signal sent by the malfunctioning sensor). The system 'decided' to push down the nose to control for the false 'belief' that the airplane goes up. So it 'decided' to crash the airplane. But, this false 'belief' was just determined by the design of the algorithmic system and the inaccurate signal sent by the malfunctioning sensor. However, the intentional stance explanations cannot use the design details to explain because the intentional stance abstracts away from any design details. So the intentional stance explanation is insufficient in this case.

In the end, it is the design of an algorithm that determines its outputs. Algorithms are designed by humans. We can design algorithms the way we like them to be. But sometimes mistakes are made. In our example, the algorithmic system is vulnerable to the inaccurate signal sent by a single malfunctioning sensor. A safer design choice would have made a difference. Had the designer anticipated that the sensor might send an inaccurate signal, she could have changed the design to make it more robust. At the very least, we should demand a design to be so that one malfunctioning sensor cannot have disastrous effects. And to figure out whether there was a malfunction, we need to scrutinize the technical details of the algorithmic system.[2] So, in at least

[2] We suspect that our example generalises: whenever an artefact is malfunctioning due to a technical detail, design level explanations are

some cases, we need to provide design explanations for algorithmic decisions.

The example illustrates that intentional stance explanations are inevitably determined by the algorithmic design. The same 'belief' of an algorithmic system might be determined by many design choices. And it matters in the Boeing case that it was this particular design which determined the algorithmic output. The specific design-level details matter to explain why an algorithm has decided falsely.

As our example indicates, transparency is important for socially consequential algorithmic decisions. These are decisions that impact people's lives and livelihoods—from loan approvals, to legal sentencing, college admissions, credit scoring, and automated transportation. We would say transparency is so important that, in some instances, algorithmic consequential decisions should be made as transparent as feasible—at least if there is no or only a reasonable cost in doing so. And Zerilli et al. agree that there is a reason to make a decision system, be it human or algorithmic, as transparent as feasible (p. 661). As a corollary, if it is feasible to make ML algorithms more transparent than humans, we have a default reason to do so.

Zerilli et al. (2019, p. 679) seek to block such a default reason to impose higher standards on algorithms by writing "to the best of our knowledge, no one has argued that algorithmic decision tools have a greater potential for transparency than human beings." But we have reason to expect that algorithms have 'a greater potential for transparency' on the design level. Due to the incredible complexity of the human brain as compared to algorithms, it is fair to posit that the architectural principles of an algorithm can, in various occasions, be more clearly specified than that of the human brain. Until we have fully understood the architectural principles of the human brain, it is plausible to take algorithms to be—in various situations—more transparent than humans. The default reason to make a decision system as transparent as feasible thus suggests a prima facie double standard of transparency.

So far, we have established that design explanations matter for at least some instances of algorithmic decision making. This suggests a double standard of transparency if we think that decisions should be as transparent as feasible. One might wonder, however, whether design explanations matter for human decisions as well. Not according to Zerilli et al. who categorically claim that the 'justification of action for human purposes takes the form of intentional stance explanation' (see quote above). Furthermore, they clearly state that "Human decision-makers [...] have never been required to furnish anything like design level explanations for their decisions." (p. 671).

Is Zerilli et al.'s (2019) dismissal of design explanations for human action justified? Well, one might argue that design explanations matter for 'defective' human behavior just like they matter for malfunctioning artefacts.[3] Consider, for example, a person who shows the symptoms of Coprolalia due to a certain neurological disorder. That is, the patient involuntarily swears at people by uttering obscene words or by making socially inappropriate and derogatory remarks. Now, would we explain the patient's behavior from the intentional stance? Rather not. The neurological condition should be cited to explain the behavior.

One could say that the patient 'decided' to swear. Having the neurological disorder can be construed as an intentional stance explanation of the swearing. We just need to be willing to treat the patient as a rational agent who has the 'desire' or 'intention' to swear at people, or the 'belief' that most people should receive the insults, or the like. After all, this would reliably predict the patient's behavior. However, we would ordinarily not ascribe the patient the intention to swear. An 'intention' to swear would not allow us to predict more than knowing that the patient has the neurological condition. We know that the patient's 'reasons' for the decision are not rational reasons and that she has no genuine desire or intention to swear (see Schroeder 2005, Sec. 3). If anything, it would be misleading to speak of a genuine decision once one knows that the patient has the disorder. And so it seems that design explanations really matter for human behaviour—contrary to what Zerilli et al. claim. But then, the standards of transparency would again 'be applied consistently across the board'. Does this mean that there is no double standard after all?

Not so quick. There is a difference of practical feasibility: while we can provide design explanations for algorithmic decisions, we cannot always do so for human actions. Neuroscience and psychology are advancing but as of yet we have no comprehensive picture of how the wiring of the human brain determines beliefs, desires, intentions, and the other mental states relevant from the intentional stance. There is still a gap in our understanding of how brain states determine mental states, partly because we have only limited epistemic access to the design of the brain. However, the architectural principles of most algorithms is more available to us as compared to that of the human brain. So we cannot always use the design of a human brain to determine the intentions of the person while we can at least try, in many cases, to use the architectural principles of an algorithm to determine its 'intentions'.

Footnote 2 (continued)

called for. Otherwise we will not understand the artefact's defective behavior.

[3] Zerilli et al. use the terms 'action' and 'behavior' interchangeably in their paper. For the purpose of this paper, we do the same.

A related point is that it is easier to change the architectural principles of an algorithm than the design of the human brain. We cannot easily change the wiring in the brain. We can, however, design algorithms the way we like them to be. But this control implies that we are more responsible for what the algorithm 'decides' than for what other human beings decide. To satisfy this responsibility we should prima facie impose higher standards on algorithms. Of course, this default might be overridden. To enact a double standard might, for instance, only threaten that powerful but opaque ML algorithms are applied wherever they could lead to a breakthrough. But it is a default nonetheless.

Pace Zerilli et al., we have argued that there are cases where design explanations should be provided and indeed perhaps preferred over intentional stance explanations. While design explanations matter in principle for both human and algorithmic decision making, it is often infeasible to give a design explanation for human decision making. And since we are responsible for the design of algorithms but not to the same degree for the design of human beings, we can and should apply higher standards to algorithmic decision making. To do justice to our responsibility is to accept the default of a double standard.

3.2 Argument from proper black boxes

The default position of a double standard is undermined if the following claim is true: an intentional stance explanation is always available save for malfunctioning systems. As a consequence, all other kinds of explanation—including design explanations—would be superfluous for functioning systems. In this section, we question this claim and its consequence.

Some have argued that we can always obtain intentional stance explanations from AI systems. Zerilli et al. (2019), for instance, write on p. 681:

> Fortunately, however, the sorts of explanations we can expect to obtain from human beings we may be able to obtain, mutatis mutandis, from AI systems too, and these really ought to satisfy the demands of explainable AI.

The 'sorts of explanations' refer to explanations from the intentional stance. If there are intentional stance explanations for AI systems, we agree with Zerilli et al. subject to the qualification in the previous section. Yet it might not always be possible to provide genuine explanations for algorithmic decisions from the intentional stance—even in absence of malfunctions.

Consider a black box algorithm, for example a certain DNN for image recognition. Let's say it has 1200 input features $x_1, \ldots, x_{1200}$. These are the lowest level features whose values represent the color of a pixel. In DNNs, the input features are combined into higher level features (computed at nodes in the hidden layers). One of these higher level features, for instance, could be the arithmetic combination $x_1^7 \cdot 1/x_2^3 \cdot x_4 \cdot \sqrt[3]{x_5}$. Sometimes higher level features can be interpreted as certain edges or color patterns. Often, however, we cannot intuitively understand what the combination of features represents. Even though the DNN might correctly classify pictures in which dogs occur from pictures in which no dogs occur (at least with high accuracy), we cannot explain what the DNN does from the intentional stance. There might simply be no most predictive interpretation under which the DNN has the 'belief' that dogs have four legs or the like. And even if a DNN had a 'belief' that is amenable to a propositional form, we might not be able to attribute this 'belief' to the algorithm. If so, we call the black box algorithm *proper*. A proper black box algorithm is ineliminably opaque in the sense that we lack epistemic access to its 'reasons' (Creel 2020). And so no genuine intentional stance explanation can be given for a proper black box algorithm.

How do Zerilli et al. (2019, p. 677) support their claim that algorithms can be given intentional stance explanations? They borrow four types of algorithmic explanations from Binns et al. (2018). The types are meant to be analogues to human intentional stance explanations, in particular the ones they call "input influence-based explanations" and "sensitivity-based" explanations. We question, however, whether these types can yield intentional stance explanations for proper black boxes.

Let us consider the input influence-based explanations. This type of explanation indicates the influence of a range of factors on the outcome. Zerilli et al. provide an example where an algorithm predicts the chances of having a car accident. Some factors are the driver's age, driving experience, and the number of trips taken at night. Now, we understand intuitively that these 'beliefs' of the algorithm correlate (some positively, some negatively) with having a car accident. So, if we understand these factors and attribute them to the algorithm as its 'beliefs', we may predict and thus explain the chances of having an accident. But does this approach also work for proper black box algorithms?

Recall the DNN algorithm for predicting the chances that a dog is in the picture. Let us assume that the feature $x_1^7 \cdot 1/x_2^3 \cdot x_4 \cdot \sqrt[3]{x_5}$ is one of the decisive factors which predicts whether or not there is a dog. It is hard for us to put the genuine content of this abstract feature into the terms of common sense folk psychology. We simply do not know how to give a genuine interpretation of $x_1^7 \cdot 1/x_2^3 \cdot x_4 \cdot \sqrt[3]{x_5}$ in mentalistic terms. And so we do not know which common sense 'belief' that p, where p is a proposition, corresponds to the abstract feature. But then, how should we describe in mentalistic terms the 'belief' whose propositional content is the abstract feature?

Sensitivity-based explanations specify factors which would need to change for the decision outcome to be different. For example, if pixel x_2 were blue instead of having its actual color, the feature $x_1^7 \cdot 1/x_2^3 \cdot x_4 \cdot \sqrt[3]{x_5}$ would change the prediction to there is a dog in the picture. But—as it is—the algorithm predicted that there is no dog. This might well be a true sensitivity-based explanation. But it is only an intentional stance explanation if we are willing to ascribe the DNN the 'belief' that the color of the pixel makes a difference as to whether or not there is a dog in the picture. Are we willing to ascribe 'beliefs' to algorithms whose propositional content we cannot epistemically access?

In the example, we do not understand the implications of certain pixels and higher-level features in mentalistic terms. Consider a new input picture where the DNN 'decides' that there is no dog in the picture. Is it still the case that the 'belief' "if pixel x_2 were blue" makes the difference? Perhaps, but not in general. There will be cases where the 'decision' is invariant with respect to the 'belief' about the color of pixel x_2. If so, the outcome of the DNN cannot be reliably predicted based on the difference-making 'belief'. This means that there will be cases of proper black boxes where we cannot reliably predict the outcomes using the intentional stance strategy. Recall that Dennett requires intentional stance explanations to successfully predict the system's future behavior—at least by and large. So, by Dennett's lights, the intentional stance does not apply here.

For proper black boxes, we are lacking epistemic access to the genuine 'beliefs' and 'reasons' of the algorithm. Sometimes we simply do not know the propositional content of a certain 'belief'. Hence, we can neither describe this 'belief' in mentalistic terms nor ascribe it to the algorithm. This problem of epistemic access carries over to the other two types of explanation (demographic-based explanations and case-based explanations). And so it remains unclear how the four types of intentional stance explanation would apply to proper black box algorithms. It seems that we have no choice for ineliminably opaque black boxes but to rely on design level explanations.

One might object that we do not need epistemic access to the genuine 'beliefs' and 'reasons' of an algorithm. Instead we only need to find an intentional stance explanation which *approximates* what the black box algorithm computes. Models and techniques for such approximations (e.g. saliency maps) are provided by the discipline of explainable AI.[4] An explainable AI system may give us an approximate intentional stance explanation for a black box but this explanation does not track what the black box algorithm actually computes. And so the approximate intentional stance explanation may not be faithful, as Rudin (2019) puts it, to the *genuine* 'reasons' of the original deep model. As the approximation does not capture the genuine reasons for the decision outcome, the approximate intentional stance explanation might not be reliable.

Adapted to our example, an explainable AI system might approximate what the abstract feature $x_1^7 \cdot 1/x_2^3 \cdot x_4 \cdot \sqrt[3]{x_5}$ means. The approximation might be expressible in mentalistic terms and so put into an intentional stance explanation. But the abstract feature is, of course, different from its approximation. And since we do not have epistemic access to the propositional content of the abstract feature, we cannot verify whether the approximate explanation faithfully captures the abstract feature. So, even if an explainable AI system provides us with an approximate intentional stance explanation, we cannot know whether the approximate explanation is faithful to the genuine 'reasons' of the black box algorithm.

Of course, the discipline of explainable AI is still in its infancy. And until we have faithful models and techniques for illuminating black box algorithms, we might not be able to obtain a genuine intentional stance explanation for a proper black box algorithm. Fortunately, however, we can still explain the decision outcomes of ineliminably opaque black boxes on the design level. So, while we wait for better systems of explainable AI, we have no choice but to impose a double standard, at least for proper black box algorithms.

We have seen that it is crucial for Zerilli et al.'s argument whether we can and are willing to attribute 'beliefs' and 'reasons' to ML algorithms. It seems to us that there is no general answer. For some transparent ML algorithms and logic-based chess computers, for example, the attribution of 'beliefs' and 'intentions' makes sense. Given such attributions, we can extensively predict their behavior and Zerilli et al.'s argument holds. By contrast, we cannot consider ineliminably opaque black boxes *as if* rational because we have no epistemic access to the 'rationale' they are following. Hence, we cannot reliably predict the decision outcomes by attributing the 'beliefs', 'desires', 'intentions', or 'reasons' which govern the decision. And so we are in no position to consider such truly opaque algorithms as rational agents or 'decision' makers. The intentional stance does not apply to proper black box algorithms.

4 Conclusion

In this paper, we have examined whether decision-making algorithms should be held to higher standards of transparency than human beings. Some scholars such as Zerilli et al. (2019) argue that a double standard is hardly justified. We have put forth two arguments for how and when a double standard is justified. First, we have argued that we need to

[4] See Guidotti et al. (2018) for a survey about the methods of explainable AI and Kasirzadeh (2021) for a critical discussion.

take design explanations into account with respect to algorithmic decision making. Second, we have made the case that the intentional stance does not apply to proper black box algorithms. The *raison d'être* of a double standard is thus supported by the need for algorithmic design explanations and the insufficieny of the intentional stance for ineliminably opaque algorithms. In this paper, we have specified two instances for which higher standards of transparency are required from algorithmic decisions as compared to humans. Based on what we have suggested, the next steps of research are a systematic exploration of the classes of algorithmic decision-making scenarios that require a higher standard of transparency, and articulation of how the algorithmic governance and regulatory proposals would look like in cases of the double standard of transparency.

Acknowledgements We would like to express our deep gratitude to Alistair Knott, James Maclaurin, and Colin Gavaghan for valuable discussions and suggestions at the University of Otago, New Zealand. Special thanks go to John Zerilli for extensive comments on an earlier draft of this paper. Furthermore, we are very grateful for the opportunity to present this work at a seminar of the Humanising Machine Intelligence project at the Australian National University. We are thankful for insightful feedback from the members of the project, in particular Seth Lazar, Sylvie Thiebaux, Damian Clifford, Pamela Robinson, and Jenny Davis. Finally, we would like to thank each other.

References

Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N (2018) 'It's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM New York, NY, USA

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, p 797–806

Creel KA (2020) Transparency in complex computational systems. Philos Sci 87(4):568–589

Davis RH, Edelman D, Gammerman A (1992) Machine-learning algorithms for credit-card applications. IMA J Manag Math 4(1):43–51

de Fine Licht K, de Fine Licht J (2020) Artificial intelligence, transparency, and public decision-making. AI Soc 1–10

Dennett DC (1987) The intentional stance. MIT Press

Feller A, Pierson E, Corbett-Davies S, Goel S (2016) A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. The Washington Post, vol 17

Gonzalez MF, Capman JF, Oswald FL, Theys ER, Tomczak DL (2019) "Where's the IO?" Artificial intelligence and machine learning in talent management systems. Pers Assess Decis 5(3):5

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv 51(5):1–42

Johnston P, Harris R (2019) The Boeing 737 MAX saga: lessons for software organizations. Softw Qual Prof 21(3):4–12

Kasirzadeh A (2021) Reasons, values, stakeholders: a philosophical framework for explainable artificial intelligence. In: Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency (FAccT 2021): 14

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215

Schroeder T (2005) Moral responsibility and tourette syndrome. Philos Phenomenol Res 71(1):106–123

Walmsley, J. (2020). Artificial intelligence and the value of transparency. AI Soc 1–11

Zerilli J, Knott A, Maclaurin J, Gavaghan C (2019) Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol 32(4):661–683