AI and Ethics
https://doi.org/10.1007/s43681-022-00217-w

ORIGINAL RESEARCH

Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque

Uwe Peters[1,2]

Received: 28 April 2022 / Accepted: 5 September 2022

Abstract
Many artificial intelligence (AI) systems currently used for decision-making are opaque, i.e., the internal factors that determine their decisions are not fully known to people due to the systems' computational complexity. In response to this problem, several researchers have argued that human decision-making is equally opaque and since simplifying, reason-giving explanations (rather than exhaustive causal accounts) of a decision are typically viewed as sufficient in the human case, the same should hold for algorithmic decision-making. Here, I contend that this argument overlooks that human decision-making is sometimes significantly more transparent and trustworthy than algorithmic decision-making. This is because when people explain their decisions by giving reasons for them, this frequently prompts those giving the reasons to govern or regulate themselves so as to think and act in ways that confirm their reason reports. AI explanation systems lack this self-regulative feature. Overlooking it when comparing algorithmic and human decision-making can result in underestimations of the transparency of human decision-making and in the development of explainable AI that may mislead people by activating generally warranted beliefs about the regulative dimension of reason-giving.

Keywords Artificial intelligence · Algorithms · Decision-making · Opacity · Mindshaping

1 Introduction

AI systems now frequently make important decisions about people including, for instance, decisions on individuals' health conditions [42], prison terms [21], or job applications [46]. The systems involved often operate according to algorithms that they did not acquire through the explicit programming by humans but by machine learning (ML), i.e., through their autonomous processing of training data and extracting correlations from them via feedback and self-correction[1] [5].

Using ML, some AI systems may discover novel correlations between certain input features (e.g., clinical symptoms) and accurate decision or prediction outputs (e.g., medical diagnoses) based on highly complex models that involve potentially millions of parameters that interact, making it hard even for AI experts to comprehend how their outputs are subsequently produced [44]. These systems are often called "black boxes" as their algorithmic decision-making (ADM) is "opaque", meaning that the internal grounds and functional processes that they act on when they produce their outputs are not fully known and an exhaustive, intelligible causal explanation of their outputs is thus unavailable [54]. While there are other AI systems, the focus here will be on black-box systems, which are primarily deep learning neural networks [34], and their ADM.

The opacity of ADM is a hot topic in the philosophy of AI and beyond (e.g., [4, 5, 11, 19, 65]). It is often invoked in the literature to question the trustworthiness of AI systems in high-stakes decision-making domains and to call for more insights into their operations or for stricter regulations on

✉ Uwe Peters
up228@cam.ac.uk

[1] Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

[2] Center for Science and Thought, University of Bonn, Bonn, Germany

[1] There are different types of ML. Three common ones are "supervised" learning (a system is trained on pre-labeled data to detect correlations and use them for classifications of new data), "unsupervised" learning (which happens without pre-labeled data and aims to detect characteristics that make data points similar to each other, producing (e.g.) clusters and assigning data to them), and "reinforcement" learning (which involves optimizing a reward function through the consequences of decisions made in interactions with the environment) [56].

Published online: 29 September 2022

their use before they are being employed in such domains [36, 53].

However, several researchers argue that this response may indicate a "double standard" or a "comparative bias" because human decision-making (HDM) is "equally opaque" [4, 65]: Empirical studies (e.g., [28]) suggest that people often do not know and cannot provide full insights into the actual basis of their decisions but instead explain them post hoc with reasons that make them appear rational. And yet, the argument continues, the trustworthiness of HDM is then not much challenged. Since that is so, it would be unfair to hold ADM systems to higher standards and require them to enable or provide exhaustive, fully accurate causal accounts of their outputs. It should be sufficient or would be advisable if these systems, too, instead provided post hoc reason-giving explanations for their decisions, or so the argument concludes (e.g., [4, 10, 65]). Call this the *equal opacity argument*. Since this argument has direct ethical implications for the debate on whether to trust ADM in high-stakes domains and on how to design and regulate the algorithms involved [25], it warrants close philosophical scrutiny.

The goal here is to critically assess the *equal opacity argument* by relating it to a particular area of philosophical research on the nature of people's ascriptions of mental states and reasons to themselves. In that research, it has been argued that when individuals ascribe mental states or reasons to themselves, this often does not involve the detection of pre-existing mental states but rather serves "mindshaping" [37], i.e., it functions to commit and govern oneself so as to think and act in ways that align with one's self-ascriptions [13, 39, 64]. While research on mindshaping has not yet been brought to bear on the debate on whether ADM and HDM are equally opaque, the claim here is that doing so is fruitful because it offers a novel and important insight for that debate.

Specifically, research on self-directed mindshaping provides grounds to believe that HDM is in some cases significantly more transparent and trustworthy than ADM. This is because, unlike AI systems that explain ADM, when human agents self-ascribe reasons to explain their decisions, their self-ascriptions frequently have a mindshaping or "regulative" function: they prompt the self-ascriber to control herself so as to conform in her thinking and acting to the reason self-ascriptions and become more predictable to other people [39, 40]. The *equal opacity argument* overlooks this point. And this oversight is problematic because it can result in underestimations of the transparency of HDM and the development of explainable AI that may mislead people by activating their generally warranted beliefs about the regulative feature of reason-giving.

My argument here is not that HDM is always more transparent or trustworthy than ADM. The claim is rather that there are empirical grounds to believe that this is sometimes the case. This is compatible with the view that ADM is overall more accurate, transparent, or trustworthy. To settle whether or not this optimistic view about ADM is correct requires a comparison between HDM and ADM after we have attained a good understanding of the full potential of both. The goal here is not to settle the HDM-ADM comparison but to help develop a fuller understanding of HDM by highlighting a fundamental difference in opacity between HDM and ADM that has gone unnoticed: the former involves self-regulation, the latter does not, and consequently the former can sometimes be more transparent and trustworthy than the latter.

Section 2 specifies the here relevant notion of opacity of ADM. Section 3 mentions recent approaches to dealing with this kind of opacity, honing in on the one commonly supported by the *equal opacity argument*. Sections 4 and 5 then introduce empirical findings cited in favor of this argument, and provide the theoretical background for a response to it by reviewing research on self-directed mindshaping. Sections 6 and 7 relate this research to ADM, and make the normative implications explicit.

2 Specifying the opacity of ADM

There are different ways in which AI systems might be said to be opaque [5, 11]. For instance, it might be difficult to understand why an AI system produces its output because the system's algorithm is too large for a human to hand-check its code line-by-line within a lifetime. This can be viewed as a kind of opacity. The focus here is on a different one.

Some AI systems are opaque in that they involve "computationally irreducible" processes such that the steps in their algorithmic operations interact in non-linear ways, involving feedback loops within the network that produce emergent properties too complex for humans to understand and trace back to particular determinative causal factors and functions ([26], p. 148–149). This is often viewed as an "in-principle opacity" because it is in these cases impossible to fully comprehend how the systems' inputs are linked to their outputs and how to disentangle the multifarious effects of multiple input interactions ([66], p. 4) even if, formally, any AI remains a "mathematical glass box", i.e., a closed system of decomposable, effectively computable operations ([33], p. 41).

For instance, when a job-recruitment black-box system decides that a particular applicant is not a suitable candidate for a vacancy, it may do so based on many different features detected on the applicant's CV that the system learned through ML to use as correct predictors for a successful application, including the applicant's work experience, GPA, health record, age, gender, and so on [61]. In some cases, a

combination of these factors or interaction(s) between the network's functional steps may lead to emergent properties that contribute to the AI's decision, creating ("structural") opacity ([11], p. 576) in that in principle no exhaustive, fully faithful, and intelligible account of the causal functions, interactions, and factors determining the AI's output can be obtained [5]. This feature of black-box systems is often linked to their accuracy such that in many cases the higher a system's accuracy in producing its outputs, the more computationally complex and so opaque its internal processing is likely to be [2].

These points about black-box systems are well known [53]. They have led to the development of ADM systems that, in addition to involving black-box AI, also include a second AI model that can provide intelligible post hoc explanations[2] of the black-box model's outputs [17, 27]. These post hoc explanations capture simplifications of a black-box system's ADM and do not contain all its technical details. Many AI explanation models treat only a subset of a black-box system's specific local decisions as explananda and seek to establish a particular feature's importance to a decision by "iteratively varying the value of that feature while holding the value of other features constant" ([58], p. 1114).

The LIME (Local Interpretable Model-agnostic Explanations) algorithm is a prominent example [52]. It explains the predictions of an AI classifier by mapping a linear model onto the patterns of its predictions, aiming to capture some of the black-box system's built-in representations such that reliable predictions can be made about its output given a particular input. Specifically, by tweaking input feature values and observing the effects on outputs, LIME learns a separate sparse linear model that locally approximates the opaque model and that enables it to extract rationales for the latter's outputs, producing easily understandable lists of most important decision factors [11]. For instance, LIME may indicate that an otherwise inscrutable AI image classifier recognizes an object in an image as an acoustic guitar because of its bridge and the fret board [6]. Relatedly, for an AI system tasked to settle whether a patient has the flu or food poisoning, LIME may produce the output 'sneeze, headache, no fatigue, so flu' [52]. These explanations resemble human reason-giving explanations of HDM and score highly on intelligibility [43, 66].

However, they have been criticized for their low fidelity and for disallowing an inspection and verification of a black-box system's decision rationale [53]. This is because these explanations do not capture how a black-box system actually forms its decisions. The algorithm whose processing (e.g.) LIME interprets does not build a sparse linear model to produce its decisions. Indeed, developers of these systems emphasize that there is typically no direct connection between the features that cause the black-box system to make its decision and those mentioned in the justification [18]. Rather, the AI explanation systems have to infer, i.e., indirectly work out, the black-box model's rationale. And indeed, as Rudin [53] puts it,[3]

> [i]f the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. […] This leads to the danger that any explanation method for a black box model can be an inaccurate representation of the original model in parts of the feature space. (p. 207)

Moreover, since the original model's processing is opaque to human agents, people cannot check the explanation's accuracy. This makes the trustworthiness of AI explanation systems and, more generally, the legitimacy of ADM questionable [14, 42].

3 Dealing with the opacity of ADM

There are different approaches. Many researchers in the field of explainable AI (XAI) focus on trying to increase the fidelity of AI explanation systems by changing their algorithmic designs (e.g., by moving from sparse linear models to generalized additive explanation models [1]). Others hold that the explanation models will remain incomplete in their fidelity and so the focus should instead be on designing models that are interpretable, i.e., not in need for an additional system that explains their processing, in the first place [54], and on developing policies to ensure AI operators state their choices, values, and operational goals in designing, building, and fielding ADM systems [32, 58].

[2] Post hoc explanations are one of several explanatory methods with which some AI systems have been equipped. For others, see [38].

[3] Rudin [53] advocates using interpretable models, not black boxes combined with AI explanation systems, because (she thinks) interpretable models are more transparent and may not always be less accurate than black boxes. She would also reject the *equal opacity argument*. I'm sympathetic to her view. But unlike Rudin, to explore the consequences, I will here grant advocates of the *equal opacity argument* that high algorithmic accuracy is inevitably linked to opacity. On Rudin's view, if we used interpretable models, we could perhaps check these models through forms of regulative reasoning not of algorithms alone but of hybrid systems comprising interpretable algorithms and human reasoners. These cases would not undermine my argument below to the effect that ADM and HDM are fundamentally different in that the latter involves a regulative dimension but not the former. This is because my argument only concerns a single HDM system compared with a single ADM system, not hybrid system cases.

The focus here will be on yet another common approach. It holds that black-box systems are less epistemically problematic than typically assumed. Different views fall within this approach (for examples see [3, 12, 16]). On the one relevant here, even if ADM is opaque and AI systems' post hoc explanations are only inferred and not completely faithful, these explanations are good enough because they correspond to the "intentional explanations" (i.e., explanations citing practical reasons including beliefs and desires) that humans themselves produce and generally view as sufficient for justifying HDM outputs [43]. In fact, many researchers hold that ADM systems *should* provide such explanations and not accounts that present the architectural innards (i.e., all the technical details) of a network's causal processing because people tend to better understand and trust them [14, 65].

For instance, Zerilli [66] argues that intentional explanations are sufficient because they "enable us to predict, at a level of accuracy significantly better than chance, what outputs a system will yield from specified inputs" (p. 9). In specifying what "at a level of accuracy significantly better than chance" amounts to when the explanandum is ADM, Zerilli proposes to "insist that the predictive accuracy of these explanations *fare no worse* than the accuracy of explanations provided by *human* decision-makers to decision subjects" ([66], p. 10). And he continues that there is "no basis for thinking that the 'reasons' of a supervised ML system would be even less faithful guides to its behavior than human reasons are to human behaviour" (ibid). This approach thus rests partly on the assumption of the *equal opacity argument*.

4 The equal opacity argument

It has been argued that requiring that ADM systems provide exhaustive, fully faithful, and intelligible insights into how they produce their outputs is overly demanding because human decision-makers, too, are equally unable to (gain and) offer full insights into all the factors causally determining their decisions [4, 10, 65]. This argument is frequently supported with reference to psychological studies. For example, citing empirical findings, Zerilli et al. [65] hold that the.

> cognitive processes underlying human choices, especially in areas in which a crucial element of intuition, personal impression, and unarticulated hunches are driving much of the deliberation, are in fact far from transparent. […] [H]uman agents are also frequently mistaken about their real (internal) motivations and processing logic, a fact that is often obscured by the ability of human decision-makers to invent *post hoc* rationalizations. (pp. 665, 666)

The assumption that "human decisions are comparatively more transparent than algorithmic decisions because they can be inspected at a depth to which AI is not presently amenable" is "false", Zerilli et al. ([65], p. 680) conclude. Notice the shift from the claim that humans are 'frequently mistaken' about their motivations (etc.) to the strong conclusion that it is false that HDM can be more transparent than ADM (i.e., if HDM and ADM are compared, there is no case when the former is more transparent than the latter).

Buckner [4] adopts a similar kind of strong claim, arguing that, just as the decision-making by AI systems such as deep learning neural networks (DNN), "[h]uman decision-making is also opaque": While AI post hoc explanation systems do not directly track all causally determinative factors of DNN outputs, psychology also "finds a disconnect between human rationalizations and the factors that actually caused the actions so rationalized" ([4], pp. 28, 32). Buckner mentions, for example, "choice blindness" studies by Johansson et al. [28]. In one of them, participants were asked to select the more attractive human face from two pictures shown to them. Shortly afterward, by the experimenter's sleight of hand, participants were presented with the picture they had *not* chosen and asked to justify their choice. Most participants did not notice the manipulation and sincerely provided reasons for choices that they did not actually make. This suggests that (just as AI explanation systems lack direct access to a DNN's decision factors) people, too, lack direct access to the factors that determine their decisions and need to indirectly infer them, Buckner holds. After mentioning further psychological evidence of such unwitting confabulations in explanations of HDM, he concludes that "unbiased assessments would score humans similarly to DNNs" regarding their lack of transparency ([4], p. 36).

This line of reasoning, i.e., the *equal opacity argument*, enjoys popularity in comparative research, and a version of it has even been traced back to Alan Turing ([36]).[4] However, I shall show that the *equal opacity argument* overlooks an important difference between the way people and AI systems are related to their own reason-giving explanations. To support this point, a brief discussion of human self-knowledge and mindshaping is required.

[4] Maclure [36] objects to the *equal opacity argument* by holding that HDM but not ADM is governed by institutionalized forms of social reasoning (judicial norms, social checks, etc.) that may neutralize individuals' cognitive limitations and make HDM more transparent and reliable than ADM. While Maclure is right to emphasize the social, institutional dimension of HDM, he, too, still seems to assume that psychological evidence of (e.g.) confabulations in explanations of HDM suggests that the decision-making of individual humans unsupported by social, institutional structures and that of AI systems are equally opaque (see [36], p. 8). I shall argue against this.

5 Knowledge of one's own reasons, and the impact of mindshaping

When it comes to one's own ability to identify why one has made a particular decision, this is a question about knowledge of one's own mind. For AI researchers who aim to compare HDM and ADM with respect to their opacity, it is thus important to be aware of work on human self-knowledge. Philosophers have written much on the topic [22]. This section offers a short primer to here relevant aspects of this work that I will then (in Sect. 6) relate to ADM.

5.1 Two approaches to self-knowledge

In the philosophical research, self-knowledge is generally thought to amount to an epistemic capacity to detect pre-existing mental states that, according to some theorists, involves a non-inferential process of introspection that provides direct access to one's own mental states ([23, 47]). Others argue that it relies on an inferential method ([7]), or an interpretive faculty that is no different from the capacity that we use to work out other people's mental states and that only provides indirect access to one's own mental states through inferences from one's own overt or covert (e.g., inner speech) behavior [8, 9]. These different proposals share an epistemic focus on the detection of one's own pre-existing mental states.

But several philosophers have called for a revision to this approach to self-knowledge. They have emphasized that ascriptions of mental states and reasons to oneself often do not only serve an epistemic but also a "regulative" [39, 45] or "mindshaping" function [13, 64], which can indirectly provide self-knowledge. The thought is that in ascribing a mental state to oneself, one often does not just describe what one detects in one's own mind (e.g., through introspection or interpretation), but one commits oneself to thinking and acting in ways that confirm the ascriptions and make oneself more predictable by other people [39]. Call this the *mindshaping view* of self-ascriptions.

To flesh it out, consider the self-ascription 'I believe that *p*'. On the mindshaping view, to hold that one believes that *p* involves adopting the social role of a believer-that-*p*, where this role is defined by the commitments, entitlements, and obligations to think and act in ways that one's interlocutors attribute to individuals with that mental state [64]. These commitments and obligations include, for instance, that one affirms that *p* if one is asked whether *p*, denies that not-*p*, and so. In uttering the self-ascription, one then opens oneself up to potential social sanctioning should one fail to act accordingly.

These points matter epistemically, advocates of the mindshaping view hold (for discussion, see [49]). This is because when people ensure a fit between their self-ascriptions and the acts that their self-ascribed mental states are meant to predict and explain, their self-ascriptions have "forward-looking" truth conditions: they become increasingly more accurate characterizations, as people are, during their development, trained to take on the responsibility for suiting their words to their actions and their actions to their words so that social coordination with them becomes easier ([39], p. 508). Mindshaping can thus provide a forward-looking insight into one's own mental states that need not involve a detection of pre-existing states. But is it reliable?

5.2 The mindshaping account of confabulations

Do people in fact commonly conform to the social roles related to their reported mental states and reasons? The choice-blindness studies [28] mentioned earlier seem to indicate otherwise. The findings suggest that (some) participants unwittingly confabulated reasons for a choice that they did not make. And so they did not act in line with their initial decisions.

However, notice that in these studies, participants do not actually offer any reasons for (nor explicit self-ascriptions of) their initial choice. And the findings in fact only suggest that (the relevant) participants did not notice that they were presented with an item, i.e., a picture of a face, that they had not actually chosen. The data *do* indicate that these study participants failed to detect that there was a difference or gap between their choices and their choice self-ascriptions. But crucially, the studies do *not* address the question as to whether and to what extent these participants are able and willing to "bridge this gap" by adjusting their thinking and behavior to align them with their self-ascriptions of choices and reasons ([13], p. 5).

In fact, some choice-blindness studies found that the attitudes that participants had self-ascribed and supported with reasons reliably matched their future behavior. For instance, in follow-up studies, when they were asked again to choose between the same face pictures at a later time, participants in the manipulation condition kept selecting the one they had offered reasons for choosing, not the picture that they had initially picked [29]. Similarly, Strandberg et al. [59] asked their participants to rate a particular policy statement and afterward gave some of them the false feedback that their positive rating of the statement applied to the *opposite* claim. Many participants did not notice this, incorrectly self-ascribed a choice different from the one they initially made, and (unwittingly) confabulated reasons. Interestingly, however, especially among participants who had justified their

choices with reasons, this attitude change persisted when they were asked about their choices even a week later, suggesting that reason-giving *solidified* their attitudes.

Turning from self-ascribed choices to self-ascribed reasons, the fact that social criticism will often ensue when agents do not align their thinking or acting with their reported reasons also supports the mindshaping view of self-ascriptions of *reasons*. Consider a case in which a human decision-maker *D* in a job recruitment scenario reports that she has decided to reject a job applicant *A* because (she thinks) *A*'s CV shows a lack of relevant work experience. Suppose that briefly afterward (without having considered new evidence) *D* defends claims or makes further decisions that are at odds with taking a lack of relevant work experience on an applicant's CV to be the basis for a rejection. It seems clear that people who learn about *D*'s conduct will usually not just readily update their views on her reasons for her decision-making and assume that sometimes the feature at issue is a reason for her to reject applicants and sometimes it is not. Rather, they will tend to criticize *D* for inconsistent decision-making. If reports of reasons did not involve any commitment, agency, and processes that can bring about the reports' confirmation, the existence of such criticism becomes hard to explain.

Moreover, while the just envisaged kind of social criticism does sometimes occur (affecting some individuals more than others), it is typically not pervasive within social groups. This indirectly suggests that people are unlikely to be routinely wrong about the causes of their decisions (*why* they want something) and right about the decisions (*what* they want). After all, living up to one's reason and choice reports requires thorough knowledge about one's own dispositions and if one's understanding of the causes of one's decisions and actions were generally unreliable, this should negatively affect one's ability to self-regulate, which in turn should routinely trigger social criticism (calls to be more consistent, make oneself more predictable, etc. [13], p. 6). Since the occurrence of the inconsistencies and social criticism at issue is limited, there is ground to believe that, overall, reason self-ascriptions capture causally determinative decision factors reliably, although cases of akrasia also suggest that their reliability is not perfect [60]. There is thus a mechanism involved in human reason-giving explanations of HDM that sustains and promotes these explanations' *predictive accuracy*, i.e., their accuracy in allowing us to predict what decisions a decision-maker will produce in response to specific inputs moving forward.

6 Revisiting the opacity of ADM

The considerations on self-directed mindshaping just outlined point to an important difference between ADM and HDM that has been overlooked in comparative research and challenges the *equal opacity argument*. This is because they provide a basis for holding that human reason-giving explanations of HDM can sometimes be significantly more transparent than AI systems' reason-giving explanations of ADM. The remainder will support this and then make explicit some normative implications.

6.1 The absence of mindshaping in ADM

As noted in Sect. 4, the *equal opacity argument* rests partly on the assumption that because people's explanations of HDM have been found to be subject to unwitting confabulations, HDM is (just as ADM) likely to involve a disconnection between the factors reported for explaining decisions and the factors that causally determine the decisions. Since this disconnection is a feature of both ADM and HDM, both are thought to be equally opaque (e.g., [4, 65]).

The preceding section suggest otherwise. It indicates that when human decision-makers report particular reasons for their decisions, even if there is a disconnection between these reasons and the factors that gave rise to the decisions, in reporting these reasons, people tend to commit themselves to thinking and acting in line with their reports. To the extent that this commitment is robust and people govern their own thinking and action accordingly, the reported reasons are no longer wholly disconnected from factors causally determining their decisions. This is not because people can change the past and alter the cause of their decisions. It is because their commitment has causal force that can merge with (or contradict) that of the original, causally determinative decision factor(s) and ensures that, moving forward, the individuals' thinking and acting is consistent with the reason self-report [13]. This can turn the self-ascribed reasons into determinative, accurately reported decision factors, and so, moving forward, make the reason-reporter's HDM more transparent to other people.

Compare this with a situation in which a black-box system is the decision-maker and an AI explanation system provides a post hoc rationalization for its output. Crucially, the processing of the AI explanation system remains purely epistemic in nature: unlike in the case of human reason reports, in the case of ADM that is made intelligible by AI explanation models, there is no feedback between the rationalizing system and the decision-making system such that the former can alter the processing of the latter. AI explanation systems do not have future-directed, regulative effects on the black-box systems whose outputs they are to

rationalize. They (currently) cannot manipulate the black-box system's processing so that it aligns with the content of the rationalization.

The reason for this is not difficult to see. In many cases, such top-down interventions and feedback would defeat the purpose of employing the black-box system in the first place. After all, these systems are used because of their computational power, which generally comes with higher complexity and so opacity [2]. If the explanation systems shaped, i.e., constrained, the more complex black-box systems' processing by pre-setting the factors causally determinative of their outputs then, assuming that higher accuracy implies increased opacity, this would reduce the black-box systems' accuracy. In fact, the explanation models would then be all we need for producing the black-box models' decision output in the first place because they would already lay out the rules that the black-box AI follows in relating inputs to outputs [53], making the use of the opaque systems redundant. The absence of top-down processing that regulates black-box models in the outlined way is thus part and parcel of preserving their superior accuracy, which, in turn, is one of the very points of using them.

6.2 Revisiting the predictive accuracy of ADM and HDM

Does the argument just outlined mean that the predictive accuracy of human reason-giving explanations of HDM is overall higher than that of AI systems' reason-giving explanations of ADM? It might seem that this is not the case. For example, it may seem that AI explanation systems do not need to regulate black-box systems to produce highly predictively accurate explanations of ADM because they are epistemically superior to humans in their explanatory tasks. Indeed, some researchers have suggested that AI explanation systems may be more accurate than humans with their accounts of HDM because, unlike human minds, black-box AI systems do not have deceptive motivations, and their processing is (still) much less complex than that of the human brain [66].

However, there are several points to note. First, as advocates of the mindshaping view have argued, it is precisely one of the key developmental and evolutionary functions of the socially regulative feature of human ascriptions of beliefs, desires, intentions, etc. and reason-giving explanations of HDM to mitigate the potential threat to the reliability of behavior interpretations and explanations of HDM posed by (e.g.) deceptive motivations and computational complexity [41, 63, 64]. Correspondingly, even if black-box systems do not have deceptive motivations and are less complex than human brains, this does not mean that AI explanations of their processing are likely more predictively accurate than human reason-giving explanations of HDM because human reason-giving explanations are also supported by mindshaping that makes people more predictable.

Second, one might have the *intuition* that AI rationalizing explanations are more predictively accurate because black-box systems cannot deceive and are not as complex as human brains. But it seems equally intuitively plausible to hold that exactly because human brains are more complex and human cognition is more sophisticated, human explanations of HDM are much more accurate. As of now, to the best of my knowledge, there is no empirical study that has pitted the two kinds of explanations against each other to compare their predictive accuracy. There is thus currently no empirical test to adjudicate between these two intuitions.

But notice that the regulative dimension of human reason-giving explanations highlighted above does provide good independent grounds to hold that these explanations are often significantly more predictively accurate than AI explanations of ADM in the following sense. Even though self-regulation, promoted by social norms, upbringing, education, etc., does not guarantee perfect predictive accuracy of human reason-giving explanations (e.g., due to akrasia), it does have the crucial feature of helping agents to *correct* predictive inaccuracy that will in many cases persist with AI systems' explanations of ADM.

To make this explicit, consider as an example a black-box system used for predicting recidivism (i.e., a convict's re-offending following prison release) [53]. Suppose that criminal history, age, and ethnicity are all correlated in the existing records. A relatively accurate explanation model of the black-box system may then output 'This person is likely to be detained because of their age' even if the opaque system does not in fact use age but ethnicity as a key decision factor. Given the mentioned correlations, the AI explanations would retain some predictive accuracy in that we could use them to produce reliable conclusions about how the system will respond (i.e., whether it will predict that a given individual will be arrested). However, since the system does not use age, the AI's explanation 'This model predicts that person X will likely be detained based on X's age' would be inaccurate, and it would stay so (unless human programmers detect and correct the error). That is, the explanation would not capture what the opaque model is actually doing and so there would be many cases in which the generalizations about how the black-box system will respond turn out to be false, as age is not a causally determinative factor in its ADM.

In contrast, if a human agent produced the kind of decision and reason report at hand, regulative agency can ensure that the reported decision factor does subsequently become partly causally determinative in future HDM: to the extent that self-regulation is effective, agents will tend to make it the case that they think and act based on the specific reasons they report. There is thus a crucial difference between human reason-giving explanations of HDM and AI reason-giving

explanations of ADM regarding their predictive accuracy. Initial predictive inaccuracy tied to the former kind of explanations can become corrected over time because of top-down control by the explaining system but not that of the latter kind of explanations, which (by assumption) are produced by systems that are causally inert with respect to the decision-making models. Thus, to the extent that human reason-givers live up to the social expectations of aligning their thinking and behavior with their reason-giving explanations, there is ground to hold that human reason-giving explanations of HDM can be significantly more predictively accurate and transparent than AI explanations of ADM.

This is not to say that human explanations of HDM are always more predictively accurate. As emphasized, akrasia and other factors may interfere with a human reason-giver's commitments to conform to their reason reports. The point here is that given their regulative dimension, human reason-giving explanations can be and, especially in HDM domains where social cross-validation is strong (e.g., in medical or judicial decision-making), likely often are more transparent than AI systems' explanations. This is because cases such as the one in the recidivism example—in which AI explanation systems track and cite excluded features, not only proxies that actually determined a model's prediction—are known to be common in explanations of opaque ML systems' outputs [53]. In these cases, the equivalent human reason-giving explanations can be more transparent to decision subjects because these subjects can count on human (but not AI) reason-givers to make efforts to guide their own thinking and acting to conform to their reported reasons moving forward. And human reason-givers, in turn, frequently succeed in their efforts and so become predictable, as indicated by the fact that charges of inconsistency and disbelief in people's reason-giving explanations are rare. Since AI explanations of ADM involve a disconnection between the reported decision factors and the causally determinative ones that is in human explanations of HDM often bridged via mindshaping, HDM is not as opaque as ADM. The *equal opacity argument*, which involves the claim that there is *no* case in which HDM is more transparent than ADM, can thus be rejected, as the points made above provide reasons to believe that there sometimes are such cases.

6.3 The opaque structures underlying HDM can have transparent sources

There is another important difference between the opacity found in ADM and HDM that is related to mindshaping and can make HDM less ethically and epistemically problematic than ADM. Notice first that the opacity found in ADM and HDM is generally thought to be problematic because it is assumed that opaque (i.e., unconscious, only indirectly accessible) processing is not based on critical reflection and a proper weighing of reasons but on uncontrolled, potentially irrelevant factors (e.g., biases [33]). This assumption may indeed seem hard to deny. However, I shall now argue that while this assumption is correct when it comes to ADM, the opacity of a human mental process such a HDM does not exclude it from also being a rational process based on conscious reasons.

To unpack this, the opaque parts of the mind that determine HDM outcomes include intuitions, fast automatic response tendencies, and heuristics. They are commonly (in psychology) referred to as "System 1" processes, which are contrasted with deliberate, slow, and conscious operations, aka "System 2" processes ([30]). Many researchers have argued that the former partly result from "habitualization", which consists in a "migration" of controlled and effortful cognitive processes into an individual's effortless, perception-like intuitive system ([62], p. 843). As Kahneman and Fredrick [31] put it, "complex cognitive operations eventually migrate from System 2 to System 1 as proficiency and skill are acquired" (p. 51).

Applying this to human expert domains, during socialization, education, and professional training as, for instance, a medical expert, one learns about the rational grounds and epistemic and social norms governing particular instances of System 2 HDM in medical domains. One then acts consistently accordingly and, through repetition, reinforcement, selective exposure, social criticism, etc., *internalizes* the relevant social norms to facilitate norm-consistent System 1 processing [51]. In this way, prior controlled processes and explicit or tacit socially endorsed norms can partly determine which fast, unconscious, and automatic intuitions emerge and how HDM will eventually proceed [50].

To illustrate this with respect to automatic unconscious cognition more generally, through repeated exposure to, for instance, counter-stereotypical information [55], people can "educate" their intuitive, often stereotype-driven System 1 judgments and restructure their opaque cognition to make it conform to rational (e.g.) moral norms ([57], p. 268). Similarly, through conscious "shifts in cognitive appraisal" ([50], p. 194), or the use of "implementation intentions" (i.e., action plans of the form 'if X, then I will do Y') to automatically react in certain ways to certain stimuli, people can produce "instant habits" ([24], p. 499) and exert top-down control, thus determining the operation and kinds of output produced by their System 1 operations. These strategies have been found to be effective (some more than others [20]), indicating that opaque HDM processing can be based on critical reflection and proper evaluation of reasons.

Returning to expert decision-making, for instance, in medical or judicial contexts, human experts' habits, intuitions, and dispositions that form the opaque structures of their HDM do not only often result from internalizing processes that were socially validated in critical reflection and

proper weighing of reasons and then encountered during professional (e.g., medical or judicial) training. They may subsequently also lead to situations in which an expert is certain about a particular professional decision but has forgotten and can no longer introspect and articulate the relevant reasons [35]. This would be an instance of opacity perhaps familiar enough to anyone who ever gradually acquired a skill through explicit instructions (e.g., doing maths). How does the ethical and epistemic status of this kind of opacity fare if it is compared with that of the opacity found in ADM?

Consider a situation in which an expert ADM system produces the same professional decision just envisaged and the AI explanation model involved also cannot provide a reason for it. On the face of it, we would be dealing with the same kind of opacity as before, and it might seem equally ethically and epistemically problematic in both cases. Importantly, however, the ADM system's opaque processing is not based on internalized, socially endorsed responses grounded in critical reflection and proper evaluation of reasons. It is based on ML in which an AI trains itself on vast amounts of data to find correlations between inputs and desired outputs on its own. In many cases, no social structures preset the ML parameters within which the AI operates [5]. This marks a key difference between ADM and HDM: to the extent that the opaque operations of HDM run within the tracks of internalized, initially reflective and socially validated structures, these operations can be less ethically and epistemically problematic than the corresponding ADM because the opaque HDM structures involved are then based on rational evaluation and reasons. This does not hold for the opaque ADM structures.

To clarify, the internalization of external structures may also produce unconscious processes that make HDM *more problematic* than ADM. This is because pernicious stereotypes, tacit discriminatory practices, unjust social inequalities, etc. may also become internalized and subsequently bias expert HDM. The thought here is not that opaque HDM structures are always based on critical reflection and a weighing of reasons, but that they can be and often are, for instance, when experts follow professional norms including those to keep biases in check. Opaque ADM structures have a different source, namely largely autonomous ML. This changes the nature of the opacity of HDM compared to that of ADM and can reduce the potential ethical and epistemic risks related to the former.

7 Normative implications

The preceding discussion has several ethically relevant upshots. One implication of the outlined differences between HDM and ADM is that human reason-giving explanations of HDM and experts' professional decisions can be viewed as in some cases more trustworthy especially in domains where social feedback on individuals' HDM is common and promotes the regulative impact of the explanations. This is because (as noted) in these cases the reason-giving explanations involve an efficacious *self-corrective* element that is missing in AI explanations of ADM and originally transparent and validated processes support human experts' opaque HDM.

There is another, related normative implication. Recall that many researchers hold that ADM systems should be designed such that they provide the same kind of reason-giving explanations that humans offer [14, 43, 65]. The preceding arguments provide grounds to believe that this view comes with risks for human decision subjects that are unappreciated so far. Notice first that designed in the way just mentioned, AI explanation systems would produce explanations that, during their development, people have learned to accept as justifications for decisions. For instance, people learn that a GP's decision to prescribe a flu medication is justified by her explaining that the patient sneezes, has headache, and so on. People partly accept this as an explanation because they assume that the GP is taking on the social role of someone with the reported reasons and so will govern herself so as to think and act in the ways that someone with these reasons would do [64]. That people indeed make this assumption is supported by the fact that social criticism will commonly ensue if the reason-giver subsequently acts in ways inconsistent with their reported reason(s). Such criticism suggests that reason-givers are assumed to govern themselves accordingly. This assumption is typically warranted because the reason-givers that people usually deal with *do* generally govern themselves accordingly or at least have the ability to do so.

Problems arise, however, when AI explanation systems are the agents that offer the explanations and there is no indication that the normative link that typically ensures the explanations' reliability via regulative mechanisms is absent. In these cases, it is only to be expected that people who receive reason-giving explanations from such agents will, just as in the case of human explanations of HDM, display a default disposition to trust them. In fact, it is precisely because people are more trusting toward and expect these kinds of explanations that many AI researchers are recommending that explainable AI be designed to produce them ([14, 15, 43]. The problem that has gone unnoticed is that if people do partly trust AI systems' explanations more because they implicitly assume that the regulative feature known from human explanations is present then their trust allocation is partly unwarranted. For these systems currently cannot play the social role that their explanations suggest that they are taking up. They cannot align

their processing with their reason reports in the way human reason-giver can do it. The proposal to design ADM systems that produce reason-giving explanations may thus lead to the development of AI systems that create a false impression of transparency and trustworthiness by co-opting an often warranted pre-existing human tendency to assume reason-givers are self-regulators. This can be epistemically harmful to the decision subjects of ADM because it prompts them to form false, overestimating beliefs about epistemic features and capacities of ADM systems.

To avoid this potential harm, decision subjects should be made aware of the absence of regulative feedback between AI explanation systems and the black-box models whose outputs they explain. Such awareness, however, is also likely to reduce people's tendency to perceive ADM systems' explanations as trustworthy because it undercuts part of the basis for their allocation of trust to them in the first place. This can be problematic because opaque AI systems are often highly accurate and thus hold great potential for people to maximize their formation of true beliefs about the world. If people's trust in these systems is undermined, this might itself create significant epistemic costs (i.e., failures to accept true propositions). There may therefore be a need for a significant normative trade-off that should be explored in future research on the risks posed by explainable AI.

8 Conclusion

Some ADM systems are opaque due to their computational complexity. This raises questions about their trustworthiness. Several AI theorists and philosophers have responded that HDM is equally opaque and so we should design ADM systems such that they provide the same kind of explanations for decisions that humans provide for HDM. Here, I challenged this argument by relating it to research on mindshaping. This research provides grounds to believe that HDM can often be significantly more transparent and trustworthy than ADM because when human decision-makers self-ascribe reasons, regulative effects occur that promote conformity with the self-ascriptions. ADM systems lack such effects. Furthermore, due to mindshaping, opaque HDM structures can be based on critical reflection and reasons, meaning that trust in opaque HDM may often be more warranted than trust in opaque ADM. Relatedly, while many researchers hold that ADM systems should be designed so as to provide rationalizing reasons for their decisions as this can increase people trust in their outputs, this can result in the development of deceptive AI. Because when AI systems provide such explanations, unlike in the case of explanations of HDM, there is no regulative mechanism that can contribute to a correction of their potential inaccuracy. Yet, people are likely to implicitly assume so. Finally and more generally, if the regulative function of human explanations of HDM is overlooked, there is a high risk that the transparency of HDM is significantly underestimated, and one might be led to the mistaken view that people's stronger trust in HDM than in ADM is unwarranted. This may incline AI developers, researchers, or the public to more willingly accept opacity in ADM even when efforts to tackle it are justified. The arguments developed here should thus be taken into account in comparative research on the transparency of AI and human cognition.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose and no competing interests to declare.

Ethical approval No ethics approval was needed.

Consent for publication Yes.

References

1. Abdul, A., von der Weth, Kankanhalli, M. & Lim, B.: COGAM: measuring and moderating cognitive load in machine learning model explanations. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–14 (2020)
2. Abdullah, T.A.A., Zahid, M.S.M., Ali, W.: A review of interpretable ML in healthcare: taxonomy, applications, challenges, and future directions. Symmetry **13**(12), 2439 (2021)
3. Asan, O., Bayrak, A.E., Choudhury, A.: Artificial intelligence and human trust in healthcare: focus on clinicians. J. Med. Internet Res. **22**(6), e15154 (2020)
4. Buckner, C.: Black boxes or unflattering mirrors? comparative bias in the science of machine behaviour. Br J Philos Sci. URL: https://www.journals.uchicago.edu/. https://doi.org/10.1086/714960 (2021). Accessed 7 Jan 2021
5. Burrell, J.: How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. (2016). https://doi.org/10.1177/2053951715622512

6. Burt, A.: The AI transparency paradox. Harvard Business review. https://www.hbr.org/2019/12/the-ai-transparency-paradox (2019). Accessed 12 Aug 2020
7. Byrne, A.: Transparency and self-knowledge. Oxford: Oxford University Press (2018)
8. Carruthers, P.: The opacity of mind: an integrative theory of self-knowledge. OUP, Oxford (2011)
9. Cassam, Q.: Self-knowledge for humans. OUP, Oxford (2014)
10. Chiao, V.: Transparency at sentencing: are human judges more transparent than algorithms? In: Ryberg, J., Roberts, J.V. (eds.) Sentencing and artificial intelligence. Oxford University Press, Oxford (2022)
11. Creel, K.A.: Transparency in complex computational systems. Philos. Sci. **87**(4), 568–589 (2020)
12. Cutillo, C.M., Sharma, K.R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K.D., MI in Healthcare Workshop Working Group: Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digital Med. **3**, 47 (2020). https://doi.org/10.1038/s41746-020-0254-2
13. De Bruin, L., Strijbos, D.: Does confabulation pose a threat to first-person authority? Mindshaping, self-regulation and the importance of self-know-how. Topoi **39**, 151–161 (2020)
14. de Fine Licht, K., de Fine Licht, J.: Artificial intelligence, transparency, and public decision-making. AI Soc. **35**, 1–10 (2020)
15. De Graaf, M. Malle, B.F.: How people explain action (and autonomous intelligent systems should too). AAAI 2017 Fall Symposium on 'AI-HRI', pp. 19-26 (2017)
16. Durán, J. M., & Jongsma, K. R.: Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics. (2021) https://doi.org/10.1136/medethics-2020-106820
17. Ehsan, U., Harrison, B., Chan, L. & Riedl, M. O.: Rationalization: a neural machine translation approach to generating natural language explanations. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, 81–87 (2018)
18. Ehsan, U., Tambwekar, P., Larry Chan, L., Harrison, B., & Riedl. M.O.: Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, 263–274 (2019)
19. Felzmann, H., Villaronga, E.F., Lutz, C., Tamò-Larrieux, A.: Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data Soc. (2019). https://doi.org/10.1177/2053951719860542
20. FitzGerald, C., Martin, A., Berner, D., & Hurst, S.: Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review. BMC Psychology, *7*, 1–12, Article 29 (2019).
21. Forrest, K.B.: When machines can be judge, jury, and executioner: justice in the age of artificial intelligence. World Scientific Publishing Company, Singapore (2021)
22. Gertler, B.: Self-knowledge. The Stanford Encyclopedia of Philosophy (Winter 2021 Edition), Edward N. Zalta (ed), URL = <https://plato.stanford.edu/archives/win2021/entries/self-knowledge/> (2021). Accessed 9 Feb 2022
23. Goldman, A.: Simulating Minds. Oxford: Oxford University Press (2009)
24. Gollwitzer, P.: Implementation intentions: strong effects of simple plans. Am. Psychol. **54**(7), 493–503 (1999)
25. Günther, M., Kasirzadeh, A.: Algorithmic and human decision making: for a double standard of transparency. AI Soc. **37**, 375–381 (2022)
26. Humphreys, P.: Extending ourselves: computational science, empiricism, and scientific method. Oxford University Press, Oxford (2004)
27. Jain, S., Wiegreffe, S., Pinter, Y., & Wallace, B.C.: Learning to faithfully rationalize by construction. ACL. https://aclanthology.org/2020.acl-main.409.pdf (2020)
28. Johansson, P., Hall, L., Sikström, S., Tärning, B., Lind, A.: How something can be said about telling more than we can know. Conscious. Cogn. **15**, 673–692 (2006)
29. Johansson, P., Hall, L., Sikström, S.: From change blindness to choice blindness. Psychologia **51**, 142–155 (2008)
30. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
31. Kahneman, D., Frederick, S.: Representativeness revisited: attribute substitution in intuitive judgment. In: Gilovich, T., Griffin, D., Kahneman, D. (eds.) Heuristics and Biases: The Psychology of Intuitive Judgment, pp. 49–81 (2002)
32. Kroll, J.: The fallacy of inscrutability. Philos Transact R Soc Part A **376**, 20180084 (2018). https://doi.org/10.1098/rsta.2018.0084
33. Leslie, D.: Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. Alan Turing Instit (2019). https://doi.org/10.5281/zenodo.3240529
34. Liao, Q. V., M. Singh, Y. Zhang, and R. Bellamy.: Introduction to explainable AI. In: Extended abstracts of the 2021 CHI conference on human factors in computing systems, pp. 1–3 (2021)
35. London, A.J.: Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent. Rep. **49**(1), 15–21 (2019)
36. Maclure, J.: AI, explainability and public reason: the argument from the limitations of the human mind. Mind. Mach. **31**, 421–438 (2021)
37. Mameli, M.: Mindreading, mindshaping, and evolution. Biol. Philos. **16**(5), 597–628 (2001)
38. Markus, A., Kors, J., Rijnbeek, P.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inform. **113**, 103655 (2021). https://doi.org/10.1016/j.jbi.2020.103655
39. McGeer, V.: Is 'self-knowledge' an empirical problem? Renegotiating the space of philosophical explanation. J Philos **93**, 483–515 (1996)
40. McGeer, V.: The regulative dimension of folk psychology. In: Hutto, D.D., Ratcliffe, M. (eds.) Folk psychology re-assessed, pp. 137–156. Springer, New York (2007)
41. McGeer, V.: The moral development of first-person authority. Eur. J. Philos. **16**(1), 81–108 (2008)
42. McKinney, S.M., Sieniek, M., Godbole, V., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**, 89–94 (2020)
43. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
44. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**, 1–21 (2016)
45. Moran, R.: Authority and estrangement. Princeton University Press, Princeton (2001)
46. Murad, A.: The computers rejecting your job application. BBC. https://www.bbc.com/news/business-55932977 (2021). Accessed 14 Nov 2021
47. Nichols, S., & Stich, S.P.: Mindreading. Oxford: Oxford University Press (2003)
48. Papenmeier, A., Englebienne, G., & Seifert, C.: How model accuracy and explanation fidelity influence user trust. http://arxiv.org/abs/1907.12652 (2019). Accessed 10 Aug 2020
49. Peters, U.: The complementarity of mindshaping and mindreading. Phenomenol. Cogn. Sci. **18**, 533–549 (2019)

50. Pizarro, D.A., Bloom, P.: The intelligence of the moral intuitions: comment on Haidt (2001). Psychol. Rev. **110**(1), 193–196 (2003)
51. Potthoff, S., Rasul, O., Sniehotta, F.F., Marques, M., Beyer, F., Thomson, R., Avery, L., Presseau, J.: The relationship between habit and healthcare professional behaviour in clinical practice: a systematic review and meta-analysis. Health Psychol. Rev. **13**(1), 73–90 (2019)
52. Ribeiro, M.T., Singh, S., & Guestrin, C.: Why should I trust you?': Explaining the predictions of any classifier. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135– 44. New York: Association for Computing Machinery. (2016)
53. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215 (2019)
54. Rudin, C., Radin, J.: Why are we using black box models in AI when we don't need to? A lesson from an explainable AI Competition. Harvard Data Sci. Rev. (2019). https://doi.org/10.1162/99608f92.5a8a3a3d
55. Rudman, L.A., Ashmore, R.D., Gary, M.L.: 'Unlearning' automatic biases: the malleability of implicit prejudice and stereotypes. J. Pers. Soc. Psychol. **81**(5), 856–868 (2001)
56. Sarker, I.H.: Machine learning: algorithms, real-world applications and research directions. SN Comput. Sci. **2**(3), 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
57. Sauer, H.: Educated intuitions automaticity and rationality in moral judgement. Philos. Explorations **15**(3), 255–275 (2012)
58. Selbst, A., Barocas, S.: The intuitive appeal of explainable machines. Fordham Law Rev. **87**(3), 1085–1139 (2018)
59. Strandberg, T., Sivén, D., Hall, L., Johansson, P., Pärnamets, P.: False beliefs and confabulation can lead to lasting changes in political attitudes. J. Exp. Psychol. Gen. **147**(9), 1382–1399 (2018)
60. Stroud, S. & Svirsky, L.:. Weakness of Will. The Stanford Encyclopedia of philosophy, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2021/entries/weakness-will/ (2019). Accessed 24 Oct 2020
61. Tambe, P., Cappelli, P., Yakubovich, V.: Artificial intelligence in human resources management: challenges and a path forward. Calif. Manage. Rev. **61**(4), 15–42 (2019)
62. Wood, W., Neal, D.T.: A new look at habits and the habit-goal interface. Psychol. Rev. **114**(4), 843–863 (2007)
63. Zawidzki, T.W.: The function of folk psychology: mind reading or mind shaping? Philos. Explor. **11**(3), 193–210 (2008)
64. Zawidzki, T.W.: Mindshaping and self-interpretation. The routledge handbook of philosophy of the social mind. Routledge, New York (2017)
65. Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C.:. Transparency in algorithmic and human decision-making: is there a double standard? Philos. Technol. **32**(4), 661–683 (2019)
66. Zerilli, J.: Explaining machine learning decisions. Philos. Sci. **89**, 1–19 (2022)