Minds and Machines
https://doi.org/10.1007/s11023-021-09570-x

GENERAL ARTICLE

AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind

Jocelyn Maclure[1]

Received: 23 February 2021 / Accepted: 16 August 2021

Abstract

Machine learning-based AI algorithms lack transparency. In this article, I offer an interpretation of AI's explainability problem and highlight its ethical saliency. I try to make the case for the legal enforcement of a strong explainability requirement: human organizations which decide to automate decision-making should be legally obliged to demonstrate the capacity to explain and justify the algorithmic decisions that have an impact on the wellbeing, rights, and opportunities of those affected by the decisions. This legal duty can be derived from the demands of Rawlsian public reason. In the second part of the paper, I try to show that the argument from the limitations of human cognition fails to get AI off the hook of public reason. Against a growing trend in AI ethics, my main argument is that the analogy between human minds and artificial neural networks fails because it suffers from an atomistic bias which makes it blind to the social and institutional dimension of human reasoning processes. I suggest that developing interpretive AI algorithms is not the only possible answer to the explainability problem; social and institutional answers are also available and in many cases more trustworthy than techno-scientific ones.

Keywords Artificial intelligence · Machine learning · Explainability · Public reason · AI ethics · Cognitive biases

1 Introduction

It is widely recognized that the deployment of machine learning-based artificial intelligence systems in all spheres of human life brings with it a host of thorny ethical quandaries that now occupy researchers and policy makers. The potential benefits of delegating tasks hitherto accomplished by humans to AI algorithms must be

The original online version of this article was revised due to retrospective open access.

✉ Jocelyn Maclure
Jocelyn.maclure@mcgill.ca

[1] Department of Philosophy, McGill University, Montreal, QC, Canada

Published online: 23 August 2021

weighed against the various ethical risks that come with such a delegation. Automated decisions can, for instance, be systematically biased against a particular class of people (O'Neil, 2016; Chander, 2017; West et al., 2019). How to allocate responsibility for the automated decisions or acts that cause harm remains a vexed question (Floridi, 2016). AI algorithms can invade our privacy by inferring information about aspects of ourselves that we did not wish to disclose by correlating data points that are not legally considered as personal information (Wachter & Mittlestadt, 2019). Even those who endorse a nuanced and prudent view of AI's capacities and foreseeable impacts on human life believe that the new wave of automation which will be enabled by AI is likely to exacerbate existing inequalities (James, 2020).

Another widely recognized problem raised by the use of AI-based socio-technical systems is caused by their lack of transparency. As I will sketch out below, the move from rule-based "symbolic" programming to machine learning dramatically improved the success rate of several AI programs, especially in fields such as computer vision and natural language processing (LeCun et al., 2015). But increased accuracy came at the cost of opacity: understanding the reasons or causes that explain why an AI system *x* decided that *y* is the right decision or course of action is generally not possible. This is what is now called, often interchangeably, AI's "black box," "explainability," "transparency," "interpretability," or "intelligibility" problem (Burrell, 2016; Weller, 2017; Selbst & Powles, 2017; Doshi-Velez & Kim, 2017; Mittlestadt et al., 2019). Accordingly, most of the high level ethical guidelines that were drafted in the past few years include explainability or transparency among the values that should underpin the development and deployment of AI systems (Jobin et al., 2019). To take but two examples, the EU's General Data Protection Regulation (GDPR) and the Government of Canada's Directive on Automated Decision-Making refer to a right to receive "meaningful information" about automated decisions (but see the qualifications in Selbst & Powles, 2017). As things stand, this general commitment to explainability remains abstract and, as I will show below, many powerful voices within AI are pleading for a weakening—if not the abandonment—of the explainability requirement.

In this paper, I will offer an interpretation of AI's explainability problem and highlight its ethical saliency. I will then try to make the case for the legal enforcement of a strong explainability requirement: human organizations which decide to automate decision-making should be legally obliged to demonstrate the capacity to explain and justify the algorithmic decisions that have an impact on the wellbeing, rights, and opportunities of those affected by the decisions. This legal duty can be derived from the demands of Rawlsian public reason. In the second part of this paper, I will try to show why what I call the argument from the limitations of human cognition fails to get AI off the hook of public reason. Against a growing trend in AI ethics, my main argument will be that the analogy between human minds and artificial neural networks fails because it suffers from an atomistic bias which makes it blind to the social and institutional dimension of human reasoning and decision-making. Finally, I will suggest that developing interpretive AI algorithms is not the only possible answer to the explainability problem; social and institutional answers are also available and in many cases more trustworthy than techno-scientific ones.

2 Machine Learning's Explainability Problem

AI's explainability problem comes from the algorithmic architecture and functioning of the AI systems currently in vogue. The new AI revival was made possible by the conjunction of three broad causes: the continuous increase in computing power more or less in line with Moore's Law, the availability of huge sets of digital data (Big Data), and the refinement of machine learning algorithms and so-called "artificial neural networks". Explainability or opacity was not a major problem for AI when rule-based "symbolic" AI was the dominant paradigm. To put it simply, "good old-fashioned AI" was based on the hypothesis that creating artificial intelligence required designing computer programs capable of logical reasoning. Programming an AI algorithm involved writing code made of a sequence of logical rules which would specify how the information fed into the machine, translated into symbols, should be processed (Boden, 2016: chap 1). "Expert systems" developed in the 1970s and 1980s included, in addition to a set of rules of inference (the "inference engine"), a knowledge base, i.e. the substantive knowledge required by the system to carry out its function, such as answering simple questions in a natural language or to identify the molecular structure of a chemical compound.

For reasons fleshed out by perceptive philosophers such as Dreyfus (1978) and by cognitive scientists drawing on phenomenology (Varela et al., 1991; Varela, 1996), the success of symbolic AI was mainly limited to virtual and contained environments such as games and logical puzzles. The failures and limitations of symbolic AI reignited the interest in how machines can learn by themselves to accomplish different cognitive tasks and in approaches inspired by how the brain works rather than by how humans reason logically. Current machine learning models and artificial neural networks sprang out of the "connectionist" paradigm in cognitive science and AI (Russell & Norvig, 2009, p. 25). To simplify, artificial neural networks are in at least a superficial sense inspired by how neurons activate and are connected through synapses in biological brains. There are several machine learning models and types of artificial neural networks, but recent successes in fields such machine vision or natural language processing are based on the neo-connectionist paradigm.

AI's opacity problem comes from the move from symbolic AI to machine learning. The lack of transparency is the price paid for improved accuracy. Whereas sequences of logical operations—such as decision trees—could in principle be extracted from the code of a symbolic AI program, machine learning algorithms require huge datasets, complex algorithmic architectures and very large numbers of parameters. As cognitive systems, machine learning algorithms are massively inductive. Their "training" often involves processing millions of data points. A deep artificial neural network is made of a large number of "layers", themselves made of interconnected mathematical units called "nodes" (depicted as "artificial neurons"). The network is organized hierarchically and each artificial neuron processes a small fragment of the data ("features"), such as the pixels representing contours, contrasts or colours in a picture.[1] The nodes acquire a particular value or weight as they process

[1] As the authors of an influential textbook put it, "[t]he hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones. If we draw a graph showing how these

the features presented in the data, and they are connected to other nodes on the basis of the strength of the relation between different data points. This complex pathway leads to the identification of patterns—large correlations—in the dataset. A computer vision algorithm can classify accurately a bike that it didn't see in its training phase because a certain aggregation of pixels is associated with the class "bike". Machine learning's inductive pattern recognition approach allows for probabilistic generalizations.[2] This is how widely used translation and computer vision algorithms work. To take a simple example, when correlated with words such money, account, deposit, and teller, "bank" is translated in French as "banque" rather than as "rive" even in the absence of an explicit rule in the algorithm specifying the possible meanings of "bank" in English.

In addition, machine learning scientists developed a feedback and self-correction procedure called "backpropagation". Once the data is "propagated forward" through the layers, a "backpropagation" algorithm sends error signals back to the hidden layers and allows for a recalibration of the weight of the different nodes and of the interconnections so that the algorithm gradually delivers more accurate results (Russell & Norvig, 2009, pp. 746–750).

Finally, to add to this already tremendous complexity, AI scientists often draw our attention to the artisanal dimension of designing deep learning algorithms. When the algorithm is designed, programmers have to tweak some of the hyperparameters and see how it changes its performance. Although theories and equations can provide some orientation in this process of trial and error, it appears that intuition and a non fully formalizable knowhow are required in the designing phase of the algorithm (Anand et al., 2020).

This helps in understanding why the computer scientist who programmed a deep learning algorithm cannot explain every single output of the AI system. No logical pathway from the input to the output can be read off the code. From a normative standpoint, this raises an obvious problem, as the decisions or predictions made by AI systems based on deep neural networks often cannot be explained. By "explained," I don't refer to a full-fledged scientific explanation; I simply mean here that no chain of reasons explaining a particular output can be easily extracted from the innards of the machine.[3]

If AI were only used for predictions made by Netflix, Amazon, and Spotify about what we might want to watch, buy, or listen to, this lack of transparency would not be ethically troublesome. But to mention only a few examples, AI is already being used, or could soon be used, in:

Footnote 1 (continued)

concepts are built on top of each other, the graph is deep, with many layers. For this reason, we call this approach to AI deep learning." (Goodfellow et al., 2016, p. 1).

[2] "Generalization" is understood here as the capacity to process new data accurately. Since the "learning" involved is predominantly inductive and statistical—statistical regularities are detected in the data—machine learning algorithms generalize in a probabilistic fashion.

[3] For instructive discussions of the various meanings of the concept of explanation in relation with AI's explainability problem, see Graaf and Malle (2017), Miller (2019b).

- healthcare for tasks such diagnosis and treatment recommendations (Raghu et al., 2019),
- human resources for selecting who is interviewed for a job (Kim & Heo 2021),
- the judicial system for sentencing or for establishing eligibility for bail or parole (Kleinberg et al., 2017)
- public administration for establishing eligibility to social assistance programmes (Booth, 2019),
- police forces for deciding where to increase patrolling or for identifying suspects (facial recognition) (Miller 2019a; Ratnaparkhi et al., 2021)
- universities for admission in academic programmes (Newton, 2021).

This very partial list shows how we are currently in the process of automating decision making in crucial areas of social life. The decisions made by AI systems or made by humans with the assistance of AI can have a direct impact on the wellbeing, rights, and opportunities of those affected by the decisions. This is what makes AI's explainability problem such a salient ethical problem.

Moreover, AI's explainability problem can have the effect of compounding its fairness problem, as the reasons leading to a decision or a prediction are not accessible. If a black women is denied a job interview or a loan, it is by assessing the justifications that we can establish whether she was discriminated against on the basis of her skin colour or gender. But if the reasons leading to the decisions cannot be made explicit and individuated, it might not be possible to assess whether a particular person was treated fairly by an organization. Only the post hoc analysis or audit of an organization's decisions could reveal that its decisions appear to disadvantage a particular group within the society.

This allows us to gain some clarity about AI's explainability problem. At least two different explainability problems can be distinguished (Pégny & Ibnouhsein, 2018). First, given the number of data and parameters involved and the tacit knowledge required to design a deep learning algorithm, researchers can find it difficult to explain precisely why and how a given AI algorithm carries out its objective function. The explainability deficit here is techno-scientific in nature. AI designers can go a long way in explaining how and why the algorithm is performing well, but some obscurity might remain. Second, given the quantity of data required in the training phase and the inductive and pattern recognition dimensions of machine learning algorithms, it is generally impossible to extract a sequence of justifications for a given output from the inner working of the system. Differently put, AI researchers have no obvious way to translate a segment of the code into a set of justifications expressible in a natural language; I will call this the "public reason" deficit. Although AI's techno-scientific explainability problem can, in some circumstances, raise ethical concern,[4] its public reason deficit is the most ethically salient.[5]

3 AI's Public Reason Deficit

From a normative perspective, a person whose wellbeing, rights, and opportunities are affected by automated decisions need not understand how the algorithm works; she needs to know the reasons why a decision that affects her was made. Although there is a variety of conceptions of social justice, the meaning of the concept of justice is centered on the right to be treated fairly; i.e., in a non-discriminatory manner, by others or by public and private organizations. The reasons or considerations that count in favour of a decision ought to be public in two ways: (1) they need to be transparent or publicly accessible and (2) they ought to be derived from, or at least compatible with, a political conception of justice. AI needs to be brought under the authority of public reason so that automated decisions can be scrutinized and assessed. This is in line with the dual aspect of Rawls' notion of public reason. Public reason, for Rawls, "is characteristic of a democratic people: it is the reason of its citizens, of those sharing the status of equal citizenship" (Rawls, 1993, p. 213). As with any mode of reasoning, public reason is normative in the sense that it incorporates "standards of correctness" and "criteria of justification" (Ibid., p. 220). In addition to its processual and procedural aspect, public reason has a substantial dimension: its object is "the good of the public"—more precisely, "constitutional essentials" and "matters of fundamental justice"—and the reasons or justifications that citizens and lawmakers offer each other when they deliberate must be drawn from a "political conception of justice" (Ibid., p. 213). A conception of justice is "political" when citizens can endorse it from the standpoint of their reasonable conception of the good.[6] Reasons are public in a substantive sense when they can pass a limited universalization test: they can in principle be accepted or seen as legitimate by all the citizens in a political community who recognize that, given the fact of reasonable pluralism, the principles of justice that underpin basic public norms and institutions should not be grounded in their personal conception of the good (Maclure & Taylor, 2011).

[4] Establishing audit mechanisms for AI systems is increasingly (and rightly) seen as one way to regulate their use (Morley et al., 2020). Auditing an AI system necessarily requires a certain level of techno-scientific explainability.

[5] When Krishnan writes that she wants to challenge "the existence and importance of a black box problem", she refers to techno-scientific explainability: "it is not clear that obtaining information about the inner workings of algorithms will be useful" (Krishnan 2020, p. 488). Later in her paper, she acknowledges that the justification of the outputs of machine learning algorithms is a genuine ethical problem. In the same spirit, Robbins writes that "[t]he real object in need of the property of 'requiring explicability' is the result of the process—not the process itself» (Robbins 2019, p. 497).

[6] I will not attempt in this paper to justify the widening of the scope of public reason that I'm suggesting here. It suffices to say for the moment that insofar as automated decisions clearly affect the wellbeing, rights and opportunities of citizens, their relation to "matters of fundamental justice" and "constitutional essentials" (Rawls 1993: pp. 227–230) is evident and strong enough.

AI's opacity problem is an ethical problem because automated decisions often fail to meet the standards of public reason. A theory of public reason is not the only ethical standpoint from which AI's explainability problem can be construed as a normative problem, but it is a particularly efficient one. It offers good reasons to think that a decision that can have a negative impact on citizens' wellbeing, rights and opportunities ought to be scrutable and assessable. AI needs to be brought within the domain of public reason because opaque automated decisions are incompatible with the principle of democratic legitimacy—the reasons that justify a decision that diminish our wellbeing or restrict our freedoms or opportunities ought to be public and open to criticism—and because they can be unfair: i.e. incompatible with a sound political conception of justice.[7]

I'll thus take it that the case for applying the standards of public reason to some AI-based decisions is strong. I will assume in the rest of this paper that AI's public reason deficit is a serious ethical problem and that this gives us pro tanto reasons for making the delegation of decision-making to AI systems legal *only if the organization can satisfy the requirements of public reason.* In other words, this entails that the sound ethical and legal regulation of AI must include what I'll call a strong requirement of explainability.[8] Since organizations ought to be able to justify their decisions, and that justification presupposes the capacity to explain a decision, the strong explainability requirement appears to rest on solid ground. I will now discuss what is probably the most widespread and prima facie convincing counterargument to the strong explainability requirement: i.e., the argument from the limitations of human reasoning. After proposing that this counterargument fails, I will reflect on the practical implications of the strong explainability requirement.

4 Minds and Machines: The Argument From the Limitations of Human Reasoning

Making the case for applying the discipline of public reason to AI and for the explainability requirement is rather straightforward. One can hardly imagine on what ground an AI enthusiast could argue that the fact that sophisticated machine learning algorithms are black boxes is not a significant ethical, political and legal problem. The standard reply is not that opaque decision-making does not contradict the standards of public reason, but rather that deep artificial neural networks *are*

[7] Reuben Binns (2018) also draws on public reason to address the related problem of "algorithmic accountability", but the analysis and argumentation provided in the paper are more programmatic than detailed and specific.

[8] In a paper that is purportedly devoted to showing that implementing a principle of explainability would be "misguided" (Robbins 2019, p. 505), Robbins ends up arguing that "[w]hat is really desired is an explanation that would provide a human with information that could be used to determine whether the result of the algorithm was justified." (2019, p. 505) In the same spirit, despite their reservations about the principles of transparency and explainability in AI ethics, Zerilli, Knott, Maclaurin & Gavaghan write that "since one cannot appeal a decision without knowing the bases upon which it has been reached, the transparency or explainability of a decision is likewise a crucial prerequisite of democratic governance." (2019, p. 663).

not significantly more opaque than human brains/minds. Let's call this the argument from the limitations of human reasoning. Decision-making, either by human beings or machines, lacks transparency. As was abundantly shown by researchers in fields such cognitive science, social psychology, and behavioural economics, real world human agents are much less rational than imagined by either some rationalist philosophers or by rational choice theorists in the social sciences. Usually relying on Kahneman's *Thinking, Fast and Slow* (2011), they point out that flesh and blood reasoners more readily use the cognitive processes linked to what is known as "System 1" when they formulate their beliefs and make decisions. System 1 operates below the radar of conscious and rational thought. It incorporates intuition and the heuristics which allow an agent to come to a conclusion without using reason for pondering evidence, weighing pros and cons, drawing inferences from premises, etc. "System 2" refers to the cognitive processes that make conscious and rational deliberation possible. Whereas System 1 is fast, unconscious, and intuitive, System 2 is slow and laborious. More troublingly, human epistemic agents often use reason in order to rationalize and justify post hoc the conclusions they reached through the heuristics associated with System 1 (Kahneman, 2011; Haidt, 2012). From this perspective, "reason" is the fancy term used to describe the capacity for making up a story after we have made a decision through System 1; in other words, motivated reasoning would be our dominant mode of reasoning. Those whom we might controversially see as defeatist about reason are prompt to point out that human cognition in general (Systems 1 and 2) is corroded by both cognitive biases and noise which hinder rational deliberation within oneself and with others. Cognitive biases are not only at play in unconscious belief and judgement formation, but also when we think that the conclusions we reached are evidence-based and logically sound.[9] Whereas biases are illegitimate criteria of judgment which lead to systematic deviations, noise refers to irrelevant factors which lead to scattered and unpredictable judgments, such as the weather, the time of day or yesterday's sport results. Both biases and noise are sources of human errors (Kahneman et al., 2021).

As a consequence, proponents of AI are prompt to point out that human cognition is also opaque, brittle, and fallible. They rightly ask what are the relevant and significant differences between human minds and deep artificial neural networks with regard to the requirements of explainability and public reason. The argument from the analogy between human minds and computers has a long and distinguished history in AI. Alan Turing himself made it in his classic paper "Computing Machinery and Intelligence":

> The short answer to this argument is that although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intel-

[9] For two distinct naturalist and evolutionary theories of human reasoning that acknowledge its limits and imperfections without concluding that it is vain, see Sperber and Mercier's "argumentative theory of reasoning" in *The Enigma of Reason* (2017), and Tomasello, *A Natural History of Human Thinking* (2014). For realistic but non-defeatist accounts of the potentialities of human reason, see Joseph Heath, *Enlightenment 2.0* (2015); Lynch, *In Praise of Reason* (2012).

> lect. But I do not think this view can be dismissed quite so lightly. […] We too often give wrong answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines. (Turing, 1950, p. 445)

This argument from the limitations of the human mind was recently taken up by the Turing Prize corecipient Geoff Hinton, often presented as one of the masterminds of the machine learning renewal. For Hinton, given that both human reasoners and artificial neural networks are not self-transparent, AI should not have to explain itself more than humans do. In an interview, Hinton opined that it would be a "disaster" should regulators insist "that you explain how your AI system works":

> I'm an expert on trying to get the technology to work, not an expert on social policy. One place where I do have technical expertise that's relevant is [whether] regulators should insist that you can explain how your AI system works. I think that would be a complete disaster. People can't explain how they work, for most of the things they do. When you hire somebody, the decision is based on all sorts of things you can quantify, and then all sorts of gut feelings. People have no idea how they do that. If you ask them to explain their decision, you are forcing them to make up a story. Neural nets have a similar problem. When you train a neural net, it will learn a billion numbers that represent the knowledge it has extracted from the training data. If you put in an image, out comes the right decision, say, whether this was a pedestrian or not. But if you ask "Why did it think that?" well if there were any simple rules for deciding whether an image contains a pedestrian or not, it would have been a solved problem ages ago. (Simonite, 2018)

The explainability requirement's critics find great comfort in Kahneman's own bleak view of human cognition and in his corresponding enthusiasm for automated decision-making based on massive computation. In an interview with the economist Eric Brynjolfsson, Kahneman said: "in general, if you allow people to override algorithms, you lose validity because they override it too often. Also, they override on the basis of their impressions, which are biased, inaccurate, and noisy. Decisions may depend on someone's mood at the moment" (Brynjolfsson, 2018).[10]

4.1 Social Reasoning and Institutional Facts

What should we make of the argument from the limitations of human cognition? To some extent, those who draw our attention to the limitations of the human mind

[10] In a more sustained fashion, Zerilli, Knott, Maclaurin & Gavaghan write that "the human brain, too, is largely a black box" (2019, p. 666). According to them, the effect of the explainability requirement for machine learning algorithm "is to perpetuate a double standard in which machine tools must be transparent to a degree that is in some cases unattainable, in order to be considered transparent at all, while human decision-making can get by with reasons satisfying the comparatively undemanding standards of practical reason." (2018, p. 668) I will challenge the view that the standards of practical reason are undemanding in the context of ethically high stake decision-making.

are right, of course. Philosophers working on human reasoning cannot conveniently ignore the research coming from the cognitive sciences on how the finite and fallible epistemic agents that we are actually think and exercise their faculty of judgement (Bortolotti, 2014). However, as I will try to show, this normalization strategy—all known forms of cognition in our actual world are opaque and fallible—is fallacious in the context of AI's explainability problem. Accepting a non-ideal theory of human reasoning does not vindicate leniency with regard to the regulation of AI. The main reason why the argument from the limitations of human reason is flawed is that it is unduly atomistic. First, it is based upon a crude and implicit form of methodological individualism: i.e., the view that the phenomena studied by the social sciences are better grasped by focusing on the thoughts, attitudes, behaviours, or others properties of individuals (Rosenberg, 2008, p. 142). Methodological individualism might be warranted with regard to some scientific inquiries, but it is inadequate as an approach to the comparison between human-based and AI-based decision procedures. Second, an equally crude form of internalism appears to be lurking in the background of the arguments made by the critics of the explainability requirement. By "internalism", I refer to the view in the philosophy of mind according to which focusing on the facts internal to the mind/brain is sufficient to understand the mind, and in particular mental content (Wilson, 2003). In the context this paper, a view is internalist if it assumes that the defects and limitations of the average individual thinker, often revealed by experimental studies conducted in an artificial environment such as the lab, are seen as an appropriate basis for a comparison between AI programs and human minds. I will not to weigh in on the internalism/externalism debate in the philosophy of mind here. It suffices for my purposes to note that the unarticulated combination of methodological individualism and internalism at play in the argument from the limitations of the human mind makes the view unduly atomistic. It is atomistic in the sense that the significance of the social and institutional dimensions of human reasoning and cooperation is lost from sight (Taylor, 1985). A particular judge, as I will try to show below, may be influenced by noise or biases, but judicial reasoning, as an institutionalized form of social reasoning, is designed with the purpose of neutralizing to the greatest extent possible the cognitive limitations of individual judges.

Those who seek to deflate the explainability problem argue that we should not be excessively troubled by the lack of transparency of automated decision-making because humans are equally opaque when they think and judge. Moreover, humans, be they police officers, judges, or employers, often make biased decisions and discriminate against the members of certain groups. From that standpoint, the relevant question is: does AI improve upon current, human-based, decision-making? If so, given the opacity and frailty of human judgement and the persistence of certain forms of prohibited discrimination, the case for a strong explainability requirement looks like an instance of shortsightedness and technophobia.

This line of thought ignores the social or intersubjective nature of reasoning and the institutional nature of most of the decision-making procedures that have an impact on the wellbeing, rights, and opportunities of citizens. The argument from the analogy between human minds and AI builds upon an impoverished social ontology. The thoroughly social and institutional character of human decision-making processes is lost from sight.[11] As public systems of rules (Rawls, 1971, p. 55), institutions are generally designed with the teleological aim of satisfying human needs and interests (Miller, 2009; Searle, 2010), including, for instance, mitigating the shortcomings of individual human reason. Consider, for instance, the case of the judges who were, on average, more severe in their sentencing just before lunch, most probably because of cognitive fatigue and hunger. Simple rules such as having mandatory breaks and a longer suspension at lunch time can attenuate the effects of fatigue and hunger. Such institutional fixes might not fully neutralize a judge's conscious or unconscious biases or vulnerability to noise. This is why a right to appeal is included in our legal rights, that the number of judges hearing a case increases as we move up the ladder in the judicial system—, for example, one trial judge, a bench of three judges on courts of appeal, and of seven judges on the Supreme Court of Canada—and that judgements are public and open to scrutiny. Resorting to an opaque AI algorithm in the judicial system cannot be vindicated by saying that individual judges are sometimes moved by obscure and arbitrary causes and that they can be biased. The intersubjective, agonistic and institutional form of reasoning at play in the judicial system is imperfect, but its norms and procedures are designed to make it fairer (Hampshire, 2001). The same logic is discernible, with more or less success, in other social practices such as hiring, police investigation, bureaucratic decision-making, etc. In non-ideal normative theory, none of these institutions are seen as perfectly capable of neutralizing human foibles, but they can be criticized and continuously improved. The answer to the defects and shortcomings of natural reason is not to throw our hands in the air, but to find ways to make decision procedures more deliberative and transparent.[12] The assertion made by Zerilli et al. (2019, p. 668) that the "standards of practical reason" are "undemanding" in comparison with what the explainability principle would entail for AI does not survive a deeper examination of our most crucial intersubjective and institutionalized modes of practical reasoning.[13] The point, to repeat, is not that social and institutionalized

[11] See Laden (2014). For the intersubjective turn in metaethics, see Manne (2013) and Maclure (2020b).

[12] For complementary institutional approaches to practical ethics, see Thompson (1999) and Weinstock (2011). A reviewer asked if the use of ensemble methods (which produce outputs by pooling the predictions made by different learning algorithms) and multi-agent reinforcement learning algorithms (which attempts to reach consensus or equilibrium among agents who either cooperate or compete) could be seen as making automated decisions more deliberative and amenable to the demands of public reason. I see these pooling methods as ways to improve, potentially, the accuracy of AI systems by drawing on the wisdom of the (algorithmic) crowd (See also Watson 2019 on the virtues of supervised learning approaches such as lasso penalties, bagging and boosting). However, this does not provide an answer to the black box problem that I am addressing in this paper. Gains in performance are compatible with even greater obscurity. As far as I am aware, pooling methods do not, as things stand, make the computations more interpretable or explainable. Although David Watson pleads for supplementing artificial neural networks with ensemble methods, he does not claim that such hybrid architectures will make AI more explainable (2019: 433).

[13] In their critical discussion of the literature on transparent or explainable AI, Zerilli, Knott et al. brush aside the intersubjective and institutional approach advocated for here. Although I can't go into details here, their dismissal is premised on a meager view of institutionalized forms of practical reasoning. According to them, intersubjective practical reasoning under the rules constitutive of our various social institutions can hardly tame our biases and other irrational impulses (2019, p. 667, 675). They accordingly have a dim view, for instance, of the judicial system (2019, p. 667). Alternatively, I hold on to the

reasoning is flawless, but that the standards and criteria of public reason provide us with the normative resources to design social institutions with the aim of making them epistemically more robust and to criticize them when they fail us.

5 The Explainability Requirement in Practice

If I am right so far, this entails that the argument from the analogy between the human mind and artificial neural networks is wrongheaded. The AI world should accept and try to satisfy the strong explainability requirement. How can this be done? Part of the answer can be techno-scientific: AI researchers are currently hard at work on making their algorithms more explainable, intelligible or interpretable. DARPA has an "explainable AI" funding program. Nobody knows yet whether this line of research will be successful. It could be that measuring the effects of an intervention on a set of hyperparameters or on the training data could throw some light on the factors that lead the algorithm to an output, such as predicting whether someone is likely to obtain a degree or reimburse a loan. The normative philosopher's role here is to try to keep up with the research and see if significant inroads are made. Many scientists are trying to program interpretive algorithms capable of shedding some light on the results of a deep learning algorithm. On pain of infinite regress, I take it that the interpretive algorithm ought itself to be interpretable and explainable. Will rule-based AI programs or expert systems be capable of such a feat? According to a team of AI researchers, the danger of whitewashing and a posteriori rationalization looms large in the explainable AI line of research (Aïvodji et al., 2019).

Insofar as AI systems are already used or are being developed for their imminent deployment in sensitive areas such as healthcare, social robotics, finance, policing, the judicial system, human resources, etc., regulators and policy makers cannot take a wait-and-see approach. They have to decide whether existing legal frameworks are capable of regulating AI. If my argument for bringing automated decision-making within the scope of public reason is sound, regulators ought to impose an explainability requirement upon the (public and private) organizations that choose to delegate decision-making to AI programs when the wellbeing, rights, and opportunities of citizens are at play. For example, if IBM's Watson Health makes a surprising diagnosis or recommends an unconventional treatment, the medical team has the duty to explain to the patient why the diagnosis was made or the treatment recommended. This is required by the norm of informed consent. To take another example, if a financial institution uses a predictive algorithm to decide whether a person qualifies for a loan, the applicant must know that it's because of some specific facts

Footnote 13 (continued)
Kantian or cognitivist position according to which practical reason has its own conception of objectivity, and the procedures of practical reason can deliver warranted or justified normative judgements. For the similarities and differences between theoretical reason and practical reason with regard to objectivity, see Rawls (1993, pp. 110–115).

about her personal finance or credit history and not because her gender or skin colour that she was turned down.

5.1 A Social and Institutional Approach to the Explainability Problem

A possible type of answer to the explainability problem that is often overlooked—perhaps because of the technological solutionism that is arguably the default position in the tech world—is here again social and institutional rather than strictly techno-scientific. Simply put, to draw upon the "context of discovery" and "the context of justification" distinction in the philosophy of science, an organization can use opaque deep learning algorithms to discover the best available answer to a question—are there cancerous cells on this scan?, should this applicant be admitted in a given graduate program?, is this citizen eligible for social welfare allocations?—and design parallel human-centered procedures for explaining and justifying their ethically-laden decisions. Those who maintain that AI ought to meet the demands of public reason need not deny that some AI algorithms perform better, under specific conditions and for well circumscribed tasks, than human reasoners. It is well established, for instance, that noise and biases degrade the quality of judicial decisions about whether defendants ought to await trial at home or in jail and about sentencing (Kleinberg et al., 2017; Kanheman et al. 2021).

In many contexts, satisfying the demands of public reason will simply require maintaining the procedures and protocols that were already in place before the introduction of machine learning tools. In healthcare, for instance, a diagnosis, prognosis or treatment recommendation made by a deep learning algorithm should be confirmed by further medical testing or by a physician's clinical judgement.[14] In the judicial system, a judge should always be able to explain and justify a particular sentence or why bail or parole is granted or not. This is a concrete way to give substance to the vague mantra that humans ought to be "kept in the loop". In science, the fact that the discovery of new ideas, hypotheses or results can be attributed to luck or error, for instance, does not necessarily undermine their validity or fruitfulness; what matters to the community of peers is how they are scientifically justified.

Maintaining and further developing human workers' competence to assess and validate the decisions made by machine learning algorithms when necessary is also a safeguard in case of errors or malfunctioning. In particular, it is well known that machine learning algorithms struggle with outliers or "long tails" (cases or events that are significant but absent or sparsely represented in the training data) and with generalization outside of the training distribution (Lévesque, 2017, p. 88; Marcus & Davis, 2019; Smith, 2019). Because of the inductive and pattern recognition nature of the learning involved in machine learning, rare data or situations may escape the wide correlations detected by current AI systems. This is another reason why

[14] Robbins appears to hold a similar view: "For example, if a medical diagnosis algorithm used as a consideration that the patient's eyes were a very specific color, we would not immediately be able to tell if this was an acceptable reason or not. This may cause us to test the hypothesis that this specific eye color was strongly correlated with the diagnosis." (2019, pp. 510–511).

AI-enabled automation should not lead to the demise of human expertise. Moreover, as Raghu et al. argue, an algorithm's superior average performance can hide "significant heterogeneity in performance" (2019, p. 2). An algorithm can perform almost perfectly on some subtasks and more poorly than humans on other subtasks. Their conclusion is that the optimal level of triage—the allocations of tasks between human experts and algorithms—is often not full automation.

Admittedly, this social and institutional approach to AI's explainability problem seems to defeat the purpose of deploying AI in the first place. Why resort to AI if decision-making is not in the end automated? Insofar as AI systems cannot explain their output when the wellbeing, rights, and opportunities of citizens are at play, and no interpretive algorithms are developed to reliably translate the interconnections within a neural net into intelligible reasons, AI should not be used to replace human-based decision procedures by algorithms. AI can still optimize decision-making by improving the overall quality of an organization's decision-making (the "context of discovery"), but this optimization does not necessarily lead to the full automation of tasks currently done by human workers.[15] Alternatively, it could be acceptable in some contexts to automate decision-making while maintaining a human-centered decision process for the litigious or more complex cases.[16] This would be in keeping with the norm according to which citizens should have the right to ask for a human intervention during a decision-making process and to obtain "meaningful information" about an automated decision (Selbst & Powles, 2017; Treasury Board of Canada Secretariat, 2019, Appendix C).

6 Conclusion

AI systems are artefacts designed to help humans better achieve their goals. AI enthusiasts are quick to ask whether it would be ethically right to forgo the benefits ensuing from delegating tasks to AI systems on the basis of their opacity. It is sometimes right, for them, to sacrifice some transparency for the sake of improved performance. As Hinton rhetorically asked in a tweet: "Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?".

[15] Although my deflationist view about AI's capacities (Maclure 2020a) leads me to reject catastrophist thinking about job automation and the "jobless economy," AI will in all likelihood lead to the automatization of many tasks and to job losses in some sectors. The social and institutional answer to the explainability problem could contribute to reducing the pace of automation on the job market.

[16] Although Kleinberg et al. believe that the use of some machine learning algorithms (such as gradient boosting) can help reducing jail time without increasing the crime rate, they acknowledge that comparing human and automated decisions is difficult, as judges may have more preferences and objectives than the algorithm. The judge might want to reduce the crime rate, jail time and racial discrimination (2017, p. 6). One of their policy recommendations is to see predictive algorithms as tools for improving the judge's decision process (rather than full automation): "given our findings that judges mistakenly release predictably-risky defendants, one natural policy response could be a warning when such an error is about to happen" (24).

Hinton's example is misleading. If a versatile robot surgeon is one day on the market, it will first have to prove that it's sufficiently safe. If it turns out that the bot wants to take the lung out of the body to operate on it and that there is no safe way to put it back, it should not be allowed near a patient. It is hard to see how can a robot surgeon be deemed safe if its actions are unintelligible. Here again, if the engineer cannot extract from the code a sequence of reasons why the robot surgeon is behaving in a particular way, the domain experts—human surgeons—need to be able to understand and explain it. Hinton presents us with a false dilemma.

Another example often used by the opponents to the strong explainability requirement is the case of self-driving vehicles. It is often said that replacing vehicles driven by humans with self-driving ones would reduce the number of road traffic injuries. Considering the erratic behaviours of human drivers, this is easy to believe. However, in practice, self-driving cars will be introduced gradually and will coexist with human drivers, cyclists, pedestrians, etc. If, as has already happened, a self-driving car hits a cyclist, the designers ought to figure out what went wrong in the computer vision system or in any other connected device in the vehicle. If they can't, the model should not be allowed on the road.

Hence, even for cases where *prima facie* it looks as if the improved performance enabled by AI justifies relaxing the explainability requirement, it is actually not clear that it is so. Weighing the benefits and the risks must be done on a case-by-case basis, but it is hard to find cases where explainability can be given up when the rights, opportunities and wellbeing of citizens are at play. Furthermore, nothing in these examples drawn from robotics suggests that a strong explainability requirement is ill-founded with regard to algorithmic governance. When a governmental organization decides to use a machine learning algorithm to assess whether a citizen is eligible for social welfare, the decision has a tremendous impact on the wellbeing, rights, and opportunities of the claimant (Booth, 2019). The governmental organization has to meet the standards of public reason and be in a position to justify its decisions. AI enthusiasts, both within and outside the AI sector, should be hard at work at designing techno-scientific and institutional solutions to the explainability problem rather than trying to banalize it.

Acknowledgements An earlier version of this paper was presented to the philosophy department at McGill University. I wish to thank all the participants for the helpful discussion. My thanks also go to Éric Gagnon for his valuable research assistance.

Funding Research Grant from the Social Sciences and Humanities Research Council of Canada, Project L'"explicabilité" des algorithmes d'intelligence artificielle: une analyse conceptuelle et normative».

Declarations

Conflicts of interest The author declares that they have no conflict of interest.

References

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S. & Tapp, A. (2019). Fairwashing: the risk of rationalization. *arXiv*. *Proceedings of the 36th International Conference on Machine Learning*, *PMLR*, *97*, 161–170. https://arxiv.org/abs/1901.09749

Anand, K., Wang, Z., Loog, M. & Gemert, J. V. (2020). Black magic in deep learning: How human skill impacts network training. https://arxiv.org/abs/2008.05981

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology., 31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Boden, M. A. (2016). *AI. Its Nature and Future*. Oxford University Press.

Booth, R. (2019). Benefits system automation could plunge claimants deeper into poverty. *The Guardian*. https://www.theguardian.com/technology/2019/oct/14/fears-rise-in-benefits-system-automation-could-plunge-claimants-deeper-into-poverty

Bortolotti, L. (2014). *Irrationality*. Polity Press.

Brynjolfsson, E. (2018). *Where Humans Meet Machines: Intuition, Expertise and Learning*. Medium. https://medium.com/mit-initiative-on-the-digital-economy/where-humans-meet-machines-intuition-expertise-and-learning-be639f00bade

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*. https://doi.org/10.1177/2053951715622512

Chander, A. (2017). The racist algorithm?. *Michigan Law Review*, *115*(6), 1023–1045. https://repository.law.umich.edu/mlr/vol115/iss6/13

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608

Dreyfus, H. L. (1978). *What Computers Can't Do: The Limits of Artificial Intelligence*. Harper Collins.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A, 374*, 20160112. https://doi.org/10.1098/rsta.2016.0112

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Graaf, M. M. & Malle, B. (2017). how people explain action (and Autonomous Intelligent Systems Should Too). *AAAI Fall Symposia*. https://www.semanticscholar.org/paper/How-People-Explain-Action-(and-Autonomous-Systems-Graaf-Malle/22da5f6f70be46c8fbf233c51c9571f5985b69ab

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon Books.

Hampshire, S. (2001). *Justice Is Conflict*. Princeton University Press.

Heath, J. (2015). *Enlightenment 2.0*. HarperCollins Canada.

James, A. (2020). Planning for mass unemployment: precautionary basic income. In S. Matthew Liao (ed.), *Ethics of Artificial Intelligence* (pp.183–211). Oxford University Press.

Jobin, A., Ienca, M. & Vayena, E. (2019). Artificial intelligence: The global landscape of ethics guidelines. https://arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf

Kahneman, D., Sibony, O. & Sunstein. R. C. (2021). *Noise: A flaw in Human Judgement*, Little, Brown Spark.

Kahneman, D. (2011). *Thinking, fast and slow*. Anchor Canada.

Kim, J-Y. & Heo, W. (2021). Artificial intelligence video interviewing for employment: perspectives from applicants, companies, developer and academicians, *Information Technology & People,* Vol. ahead-of-print, No. ahead-of-print. https://doi.org/10.1108/ITP-04-2019-0173

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Krishnan, M. (2020). Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology, 33*, 487–502. https://doi.org/10.1007/s13347-019-00372-9

Laden, A. S. (2014). *Reasoning: A Social Picture*. Oxford University Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. https://doi.org/10.1038/nature14539

Lévesque, H. J. (2017). *Common Sense, the Turing Test, and the Quest for Real AI*. The MIT Press.

Lynch, M. P. (2012). *In praise of reason: Why rationality matters for democracy*. The MIT Press.

Maclure, J., & Taylor, C. (2011). *Secularism and freedom of conscience*. Harvard University Press.

Maclure, J. (2020a). The new AI spring: A deflationary view, *35*, 747–750. https://doi.org/10.1007/s00146-019-00912-z

Maclure, J. (2020b). Context intersubjectivism, and value: Humean constructivism revisited. *Dialogue: Canadian Philosophical Review/revue Canadienne De Philosophie, 59*(3), 377–401. https://doi.org/10.1017/S0012217320000086

Manne, K. (2013). On being social in metaethics. In R. Shafer-Landau (Eds.), *Oxford Studies in Metaethics,* Vol 8. (pp. 50–73). Oxford Scholarship Online. https://doi.org/10.1093/acprof:oso/9780199678044.001.0001

Marcus, G. & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Miller, S. (2009). *The moral foundations of social institutions: A philosophical study*. Cambridge University Press. https://doi.org/10.1017/CBO9780511818622

Miller, S. (2019a). Machine learning, ethics and law. *Australian Journal of Information Systems*. https://doi.org/10.3127/ajis.v23i0.1893

Miller, T. (2019b). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B., Russell, C. & Wachter, S. (2019). Explaining Explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19) (pp. 279–288). *Association for Computing Machinery*. https://doi.org/10.1145/3287560.3287574

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Newton, D. (2021) Artificial Intelligence grading your 'neuroticism'? Welcome to college's new frontier. *USA Today*. https://www.usatoday.com/story/news/education/2021/04/26/ai-infiltrating-college-admissions-teaching-grading/7348128002/

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Pégny, M. et Ibnouhsein, M. I. (2018). Quelle transparence pour les algorithmes d'apprentissage machine ?. *Archives-Ouvertes*. https://hal.archives-ouvertes.fr/hal-01877760

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., & Mullainathan, S. (2019). The algorithmic automation problem: prediction, triage, and human effort. https://arxiv.org/abs/1903.12220

Ratnaparkhi, T. S., Tandasi, A., & Saraswat, S. (2021). Face Detection and Recognition for Criminal Identification System, *Xplore, 11th International Conference on Cloud Computing, Data Science and Engineering (Confluence),* 773–777. https://doi.org/10.1109/Confluence51648.2021.9377205

Rawls, J. (1971). *A theory of justice*. Belknap Press of Harvard University Press.

Rawls, J. (1993). *Political liberalism*. Columbia University Press.

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds & Machines, 29*, 495–514. https://doi.org/10.1007/s11023-019-09509-3

Rosenberg, A. (2008). *Philosophy of Social Science (3rd edition)*. Westview Press.

Russell, S. & Norvig, P. (2009). *Artificial intelligence: A modern approach*. (3rd edition). Pearson.

Searle, J. (2010). *Making the social world*. Oxford University Press.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Simonite, T. (2018). *Google's AI guru wants computers to think more like brains*. WIRED. https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/

Smith, B. C. (2019). *The promise of Artificial Intelligence: Reckoning and judgment*. The MIT Press.

Taylor, C. (1985) Atomism. In C. Taylor, *Philosophy and the Human Sciences. Philosophical Papers vol. 2* (pp. 187–210). Cambridge University Press.

Thompson, D. (1999). The institutional turn in professional ethics. *Ethics & Behavior, 9*(2), 109–118. https://doi.org/10.1207/s15327019eb0902_2

Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.

Treasury Board of Canada Secretariat. (2019). *Directive on Automated Decision-Making*. Government of Canada. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 49*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. The MIT Press.

Varela, F. J. (1996). *Invitation aux sciences cognitives*. Seuil.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review, 2019*(2), 494–620. https://doi.org/10.7916/cblr.v2019i2.3424

Watson, D. (2019). The rhetoric and reality of anthropomorphism in Artificial Intelligence. *Minds and Machines, 29*(3), 417–440. https://doi.org/10.1007/s11023-019-09506-6

Weinstock, D. (2011). How political philosophers should think of health. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine, 36*(4), 424–435. https://doi.org/10.1093/jmp/jhr026

Weller, A. (2017). Transparency: Motivations and Challenges. *arXiv*. *Proc. ICML Workshop Human Interpreting Machine Learning*, 55–62. https://arxiv.org/html/1708.02666

West, S. M., Whittaker, M. and Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. *AI Now Institute*. https://ainowinstitute.org/discriminatingsystems.pdf

Wilson, R. A. (2003). Individualism. In S. Stich & T. A. Warfield (Eds.), *The blackwell guide to philosophy of mind* (pp. 256–287). Blackwell.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

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