Philosophy of Science (2022), **89**, 1–19
doi:10.1017/psa.2021.13

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

Explaining Machine Learning Decisions

John Zerilli

University of Oxford, Oxford, UK
Emails: john.zerilli@law.ox.ac.uk; john.zerilli@gmail.com

(Received 06 April 2020; revised 21 August 2020; accepted 29 October 2020)

Abstract

The operations of deep networks are widely acknowledged to be inscrutable. The growing field of Explainable AI (XAI) has emerged in direct response to this problem. However, owing to the nature of the opacity in question, XAI has been forced to prioritise interpretability at the expense of completeness, and even realism, so that its explanations are frequently interpretable without being underpinned by more comprehensive explanations faithful to the way a network computes its predictions. While this has been taken to be a shortcoming of the field of XAI, I argue that it is broadly the right approach to the problem.

1. Introduction

The operations of the most advanced machine learning (ML) systems in use today are widely acknowledged to be "opaque," "inscrutable," or "black boxes" when compared with more traditional forms of ML. The growing field of Explainable AI (XAI) has emerged in direct response to this problem. XAI can be seen to espouse three overriding aims for explanations of ML decisions: (1) completeness or depth; (2) realism/fidelity; and (3) interpretability. In XAI, complete explanations are those which purport to be exhaustive, bringing to light the architectural innards of a tool and their systematic operations. As I am using the word "deep," it will also apply to exhaustive explanations, but explanations in XAI can be deep without being exhaustive, most straightforwardly by bringing to light the operations of *parts* of a learning tool only, e.g., specific input features, parameters, or calculations (Lipton 2017). Real explanations are those that are faithful to the way a system computes its predictions (Rudin 2019; Guidotti 2018). Lastly, as I shall be using the term "interpretable," an interpretable explanation is one which can be understood by a decision subject, often through deployment of agent-level or folk-psychological categories.[1] Interpretability is a feature of explanation that is especially important when the subject of an automated decision needs to know how a particular decision regarding them was reached (e.g., for the purposes of challenge or appeal), and indeed, in this paper, I will be

[1] Occasionally in the XAI literature, the predicate "interpretable" modifies a system or its operations rather than an explanation. As I have defined it, interpretability is a property of explanation, not of a system or its operations.

concerned solely with the desiderata of explanations that purport to justify automated decisions. This is because most of the research effort in XAI is concerned with explainability as a means for enabling interested parties to assess whether an automated decision is justified (Selbst and Barocas 2018). Other goals that might be served using explainable systems, such as control over and improvement of these systems, have tended to be secondary, if still in the background (e.g., Adadi and Berrada 2018).

Even though interpretability may not be compatible with either completeness or realism, in the sense that an explanation that is interpretable will most likely not *also* be complete or realistic; still, it is generally accepted that an interpretable explanation should ideally be underpinned by (and referable to) a more complete and faithful explanation (Rudin 2019; Leslie 2019; Guidotti 2018; Lipton 2017). In fact, complete explanations are only possible for certain types of ML systems, namely, those whose operations are *fathomable* or in some other way *intelligible* to a trained expert. Because the operations of the most advanced ML systems are not fathomable, nor even readily intelligible, in many cases XAI prioritises interpretability at the expense of completeness, and sometimes even realism, so that explanations are frequently interpretable *without* being underpinned by more comprehensive explanations that are faithful to the logic of the system. While this might be—and has been—taken to be a shortcoming of the field of XAI (Rudin 2019; Leslie 2019), I argue that it is broadly the right approach to the problem. In particular, I argue that two features of action explanation are in many cases *jointly sufficient* for being able to assess whether an automated decision is justified:[2] the intentional structure of the proffered explanation (or, as I explain in section 4.5, something approximating this structure), as well as the ability the explanation affords for tracking the system's behaviour.

The paper can be understood as a philosophical defence of a significant body of work in XAI—a body of work which, whether consciously or otherwise, aims to explain ML systems by reducing their operations to a form that is amenable to belief-desire representation. I submit that this philosophical defence is both timely and worthwhile because extant criticisms of XAI explanations seem to assume that such explanations are deficient as forms of justifying explanations precisely because they lack completeness/depth and realism. That is to say, XAI explanations are considered deficient for having *only* those features of belief-desire explanations that I contend are (often) sufficient to qualify a given explanation as a justifying explanation.

The paper is organized as follows. Section 2 will briefly survey the principal ML techniques I have in mind, and show why, relative to earlier expert systems and simple linear models, they are thought to pose unique challenges for explainability. Section 3 will briefly describe Daniel Dennett's intentional systems theory, which may be understood as an analysis and (qualified) defence of folk psychology. A folk psychological model is not the only model one might adopt for interpreting human action, but as models of human action go it is unquestionably the most prevalent form. Most importantly—so far as the present discussion is concerned—*reasons for decisions*, including official and even high-stakes decisions, are generally expressed in the vernacular of folk psychology. Considering this, section 4 argues that in assessing

[2] I use the words "action," "decision," and "recommendation" interchangeably.

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

whether an automated decision is justified it will often be sufficient for its explanation to have some kind of intentional structure that is able to track the system's behaviour. Section 5 considers when explanations deeper than those offered by the intentional stance will be required. I conclude in section 6.

2. Why explanation is hard for machine learning

2.1 Types of machine learning

Machine learning is a form of data processing that identifies statistical patterns from large quantities of information. Instead of being programmed with predetermined responses to a set of conditions—the dominant approach to AI up until fairly recently—an ML system is set up to "learn" its own suitable responses to those conditions under a training regime. Many tasks for which no straightforward sequence of "if-then" rules can be formulated may be handled more efficiently, and indeed more effectively, by a system able to draw its own inferences from a database of instances.

There are two main classes of ML systems: "supervised" and "unsupervised." Generally speaking, supervised ML systems assist with prediction—involving, in the simplest case, a mapping from some known input x to an unknown output y. This mapping is made possible through training, which in turn takes place on a set of pre-labeled input-output pairs $(\mathbf{x_i}, y_i)$. If we wanted to train a system to recognise the image of a dog or a cat, for example, we might conceive of $\mathbf{x_i}$ as a feature vector (consisting of all the notable features of the object in question, e.g., ear, tail, fur, nose, etc.), and conceive of y_i as the set of class designations corresponding to those features (we can specify that y takes one value in $\{1, \ldots, C\}$, where C stands for the number of classes; here $C = 2$, i.e., DOG or CAT). In this example, "supervision" would consist of a human labeler "telling" the system which features pertain to a dog and which to a cat. If $D = \{(\mathbf{x_i}, y_i)\}$, where D is the training set, then we can say that as the size of D grows, the more "experience" the system acquires and the more accurate the system will be in its predictions when used on unseen data. The difference with unsupervised learning is that there is no similar process of labeling—no output data y_i used in training—so $D = \{\mathbf{x_i}\}$. (Unsupervised learning is thus concerned less with prediction than with description or the discovery of patterns in datasets.) My focus in this paper will be on supervised ML systems, because these are the ones being increasingly used to supplant or supplement human decision-making, particularly in areas such as criminal justice (e.g., when determining risk of flight or recidivism in bail and parole hearings), legal practice (when assessing a client's prospects of success), medicine (for diagnosing a patient's illness), finance (in assessing a loan applicant's credit-worthiness), and public administration (for passport verification, assessing welfare recipient entitlements, etc.).

2.2 The challenge posed by neural networks

Not all ML systems are thought to be opaque. In the main, it has only been neural networks which have been considered to pose serious problems for explanation. Linear and logistic regression techniques, decision trees, and even random forest forecasting, do not pose the sorts of difficulties that plague neural networks.

The opacity of an ML model is often defined by reference to the properties of *linearity* and *dimensionality*. Linearity and low dimensionality together can be said

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

to make a system and/or its operations *fathomable*, in the sense that "a person can contemplate the entire model at once" (Lipton 2017, 4). Thus, for a system to be fathomable, "a human should be able to take the input data together with the parameters of the model and in *reasonable* time step through every calculation required to produce a prediction" (Lipton 2017, 4-5).[3] But when dimensionality increases considerably, even a linear model will cease to be fathomable. Dense linear models, extensive rule lists, and large ensemble methods are really of intermediate complexity, and, in many cases, too cumbersome for a person to work through stepwise in real time without losing their footing. Nonetheless, so long as linearity is a feature of the model, it will still be *intelligible*—i.e., inspectable—without necessarily being fathomable—i.e., inspectable *all at once*.

The most complex systems, such as deep learning networks, are inscrutable to the extent that they model relationships that are *not* linear and incorporate *extremely large* feature spaces. Their operations are thus neither fathomable nor intelligible, in the above senses. There is also the added complication that their model parameters are learned quasi-independently, even in supervised learning scenarios. These characteristics combine to render neural networks opaque in a way that is categorically different from the way complex linear models may be difficult to parse. When the relationships within linear models are obscure, they are obscure by reason of the merely practical and attentional limitations posed by human cognition, such as working memory. For this reason, they may still be considered intelligible (if not fathomable). There is a sense in which the opacity of neural networks is an *in-principle* opacity, because it is practically impossible to appreciate how their inputs relate to their outputs and to disentangle the multifarious effects of multiple input interactions. This does not mean, however, that such systems are a ding an sich. From a purely formal point of view, black box systems constitute a "mathematical glass box." Even when a neural network takes into account many millions of parameters, still "an algorithmic model is a closed system of effectively computable operations . . . . In this restricted sense, all AI and machine learning models are . . . transparent . . . " (Leslie 2019, 41). I shall call this property the *tractability* of a system. It is a property that all technical systems share, from the simplest to the most complex. But the kind of epistemic access into the operations of a system that tractability affords does not permit us to explain how the system reasons through to its conclusions, and thus how it decides matters in particular cases. Put otherwise, the formal explanation of a system would not constitute a *semantic* explanation, which *does* allow us to understand "the functions of the individual parts of the algorithmic system in the generation of its output" (Leslie 2019, 41–42).

Some of the issues in play here can be clarified by reflecting on the experiences of connectionists with the first neural networks developed in the 1980s. The computational psychologist, David Marr (1977), noted a distinction between what he called "Type 1" theories and "Type 2" theories. Type 1 theories model those aspects of phenomena that yield to systematic task analysis. In later work, Marr (1982) would adumbrate a general methodology for Type 1 investigation, involving his well-known "computational," "algorithmic," and "hardware" levels. By contrast, Type 2 theories

[3] Lipton (2017) himself describes such models as "simulatable," which is an apt choice, but I prefer the epistemic connotations of "fathomable."

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

Table 1. The terrain of ML systems and the nature of the epistemic access each affords

		Epistemic access afforded		
		Tractability	Intelligibility	Fathomability
System	SIMPLE e.g., simple linear regression	X	X	X
	INTERMEDIATE e.g., regression extensions, long rule lists, boosted decision trees, random forests	X	X	
	COMPLEX/"TYPE 2" e.g., convolutional networks, support vector machines, humans	X		

model such aspects of phenomena as depend on "the simultaneous action of a considerable number of processes, *whose interaction is its own simplest description*" (Marr 1977, 38). These include high-level cognitive tasks and range anywhere from medical diagnosis to literary composition (Boden 1990). In general, Type 2 theories are less likely to be discovered than Type 1 theories, and Marr thought that if a phenomenon is too complex to yield to Type 2 explanation, the hope of its being comprehensively understood must be abandoned. Human expertise and literary appreciation are among tasks that could be considered so complex that any Type 2 explanation of them would be unintelligible (Boden 1990).

In retrospect, it can be seen that Marr's Type 2 explanations are akin to "mathematical glass box" explanations, which may be unintelligible though formally precise (see table 1). Moreover, while connectionist networks may be amenable to Type 2 explanation (e.g., in terms of synaptic weights, Boltzman equations, etc.), they do not readily submit to Type 1 task analysis (Clark 1990). I have been calling a system which yields solely to Type 2 explanation one that is merely *tractable* without also being intelligible or fathomable, but we could just as well call such systems "Type 2 systems." To repeat, these are systems that do not yield to semantic explanation: they defy most systematic attempts to comprehend them. In the 1980s, this was basically conceded. But there was, pace Marr, judged to be a way around the conundrum. The best hope of explaining Type 2 systems lay in being pragmatic about what counts as a suitable explanation. This in turn required entertaining "the explanatory role of elucidating various structural possibilities, within which natural phenomena must lie and in terms of which they can be systematically compared" (Boden 1990, 9–10). In what Andy Clark (1990, 211) called "the methodology of connectionist explanation," the a priori, logicist, and typically ad hoc axiomatisation of a domain of intelligence was eschewed. Because a network is constructed in advance of any such axiom system being devised, and the successful reproduction of intelligent behaviour is key, explanation can afford to be post hoc—abstract principles will be discovered a posteriori. But neither Type 1 nor Type 2 theories were considered appropriate. Instead, a battery of techniques was recommended as the occasion dictated: deliberate interventions on a network's architecture to see what will happen ("network pathology"); "local" or "cluster" analyses of activation patterns within specific systems; and so on.

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

As we shall see in section 4.5, this connectionist methodology does not just resemble, it positively characterises a significant family of approaches used to produce interpretable explanations in XAI today.

3. Framing human interpretability in terms of the intentional stance

Without doubt, the most prevalent and user-friendly model for interpreting human action is everyday "mentalistic," "commonsense," or "folk" psychology—an interpretation of action that traffics in the familiar beliefs, hopes, expectations, hunches, loves, wishes, desires and longings of ordinary human striving. It is the most natural idiom for humans to adopt when seeking to explain their actions to one another (as well as to themselves), and, far from being limited to merely informal settings, frames the dialogic structure of much judicial, administrative, and even commercial decision-making ("Offender X was given a lenient sentence because Judge Y believed X was remorseful, and genuinely wanted to rehabilitate," "Company X wanted to expand into this new area of the market to forestall what it believed was its competitor's attempts to take advantage of renewed customer satisfaction regarding Product Y," etc.). In short, the syntax of folk psychology is a paradigm of practical reasoning.

Perhaps the best attempt to make sense of how this kind of explanation fits in with—and fares in relation to—other systems of explanation is Daniel Dennett's intentional systems theory (Dennett 1971; 1987; 1991; 2009). In accounting for the efficacy of folk psychological explanation in practical reason, Dennett's account situates folk psychological explanation within a framework of three systems of explanation, the categories of which are only accessible by adopting one of three "stances" towards an explanandum. Folk psychological vocabulary and the efficient, user-friendly explanations it makes possible are accessible by adopting what Dennett calls the *intentional stance*: "The intentional stance is the strategy of interpreting the behaviour of an entity (person, animal, artifact, whatever) by treating it *as if* it were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires'" (2009, 339). An intentional system is, by definition, "[a]nything that is usefully and voluminously predictable from the intentional stance" (2009, 339)—in other words, any explanandum for which the intentional stance pays dividends: humans most obviously.[4]

The intentional stance packs a powerful punch, and this can be seen most clearly when it is set against the two other stances (or "strategies of prediction") Dennett describes. The *physical stance* "is simply the standard laborious method of the physical sciences in which we use whatever we know about the laws of physics and the physical constitution of the things in question to devise our prediction" (2009, 340). Mass, velocity, atoms, and molecules are the stuff of physical stance explanations. And since everything is *at least* a physical system, everything's behaviour can be predicted from this stance. The only drawback is cost. The physical stance can be a tedious affair, and "seldom practical" for anything exhibiting design, such as artifacts and living creatures. Enter, therefore, a "fancier style of prediction": that made possible by adopting

[4] In Dennett's own writings, the intentional strategy has been variously applied to mammals, birds, fish, reptiles, insects, spiders, clams, computers, and thermostats (e.g., Dennett 1987).

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

the *design stance*. From this perspective, the fine-grained detail of a system's physical constitution can be screened off, so that much more abstract, essentially functional, features now explain how the parts of a system contribute to the end in view of which the system exhibits design. In Dennett's words: "Nobody would prefer to fall back on the fundamental laws of physics to predict the behaviour of a chainsaw when there was a handy diagram of its moving parts available to consult instead" (2009, 340).

To inhabit the intentional stance is to inhabit a still more rarefied point of view in which "the designed thing is treated as an agent of sorts, with beliefs and desires and enough rationality to do what it ought to do given those beliefs and desires" (2009, 340). This posture is not warranted when an artifact is relatively simple, such as an alarm clock, but becomes "well-nigh obligatory" when the artifact approaches a certain level of complexity. Over many years (e.g., 1971; 2009), Dennett's preferred examples for making this case have always been chess-playing computers:

> just think of them as rational agents who *want* to win, and who *know* the rules and principles of chess and the positions of the pieces on the board. Instantly your problem of predicting and interpreting their behaviour is made vastly easier than it would be if you tried to use the physical or the design stance. (Dennett 1971, 340)

In this passage, Dennett helpfully cites both *predicting* and *interpreting* the computer. The one complements the other. As for prediction: knowing the computer's "beliefs," such as those regarding which moves in the game are legal, as well as its "desires," in this case winning the game (taking the king), we can predict with a fair degree of accuracy what its next move will be. Taking stock of the chessboard and noting all the legal moves the computer could take would give you a roster of the system's beliefs. In light of the computer's goal of winning the game, you could then pretty straightforwardly rank the 20 or 30 moves in the offing from wisest/most rational to least wise/most irrational. On the assumption that you are dealing with a rational agent, your prediction would be that the computer will choose perhaps any of the top-four-ranked moves. In a game setting, this counts as "tremendous predictive leverage" (2009, 341). As for interpreting the computer's move ex post—i.e., after it has made its move—a good explanation here then need only reference the system's beliefs and desires, for a "good" explanation is simply one that gives you what we need to assess the quality of the action. To be told that the computer chose to move its knight because that move ranked ahead of all other available moves (once the risks and gains attendant on all other moves are factored in) is to be offered a *justification* for the move, albeit one that is fully comprehensible only in light of the norms of competent chess-playing. More precisely, it is to be furnished with reasons for action which, in normal circumstances, will suffice to determine whether an action was justified.

No doubt the most accurate predictions would stoop to the microstructural kinds of the physical stance and entail such activities as "calculating the flow of electrons that results from pressing the computer's keys," or get on just as successfully—and *much* more easily—from the design stance, "considering the millions of lines of computer code that you can calculate will be streaming through the CPU of the computer after you make your move" (2009, 341). But both of these options, whether adopting the design stance or the physical stance, come "at a tremendous cost of time

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

and effort," with the predictive leverage thus gained worth neither the time nor the effort. The intentional stance, by contrast, affords a respectable predictive strategy that is also, by its very compendious abstractness, considerably easier to handle: the set of complex factors from which human choices actually arise are conveniently compressed and even idealized, much like the cleaner, sharper—if imperfect—patterns "dimly discernible" in a noisier bit map image (Dennett 1991). In section 4, I will argue that an intentional explanation, in virtue of these very structural and behaviour-tracking properties, satisfies two conditions that are, at least in many cases, jointly sufficient for assessing the justifiability of a decision.

4. Interpretability in XAI

4.1 Outline

I have said that in assessing whether an automated decision is justified, it will often be sufficient for its explanation to have some kind of intentional structure that is able to track the automated system's behaviour. I shall now defend this position more systematically. The argument moves in four steps: first, I defend the claim as a consequence of the fact that automated decision systems function in loco hominum (i.e., in the place of a human), and since these features of explanation are generally taken to suffice for explanations of human decisions, prima facie they ought to suffice for explanations of machine learning decisions; second, I argue for the sufficiency of these features of explanation *in their own right*, i.e., without regard to their salience in human practical reasoning; third, I defend the adoption of an intentional stance towards ML systems; finally, I illustrate through various examples from XAI how explanations of ML decisions might be intentionally structured and track system behaviour.

4.2 Parity of humans and machines in loco hominum

So long as we install machines in human offices (in loco hominum), there seems to be a prima facie case for holding them to the same norms governing human occupancy of such offices. This position seems especially reasonable when exercise of the decision-making function that an ML system is given necessarily affects the rights or interests of a human subject. Thus, if intentional explanations are considered good enough for assessing human decisions, they should be considered prima facie good enough for assessing automated decisions. This consideration is defeasible, as there may occasionally be competing considerations that outweigh it, perhaps arising from features specific to certain automated systems (see e.g., section 5.2). But as a prima facie rule, it seems plausible.

However, it is not just the fact that machines are placed in loco hominum. It is significant too, that a certain *kind* of machine is being so placed. The automated systems we have been contemplating are Type 2 systems. But then the human practical reasoning system is also a Type 2 system; indeed, it was the original Type 2 system from which earlier connectionist models merely derived their Type 2 status (connectionist systems were, after all, loosely inspired by the microanatomy of the human brain). Absent countervailing considerations (see section 5), it makes little sense to impose radically different standards of explanation on systems that afford the same kind of epistemic access—systems, in other words, which have parity not

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

only in respect of how they are situated, as decision agents, but also in respect of the epistemic access they afford. And one way that their Type 2 status manifests is in the somewhat eclectic (and satisficing) "methodology of connectionist explanation" that provides the most promising avenue for understanding them (see section 2.2).

4.3 Brevity and predictive accuracy

Intentional structure and behaviour-tracking, quite apart from their salience in explanations of human action, can be jointly sufficient in accounting for action *in their own right*.

First, intentional structure confers *brevity*, or "maximum information with the least cognitive effort," which in turn reduces the explanandum to "behaviourally and cognitively usable proportions" (Rosch 1978, 28). A municipal authority that resolves to ban the construction of buildings above a certain height *could* (let us imagine) require each of its voting members to have their brains scanned, with a view to producing high-fidelity images of the state of each voting member's brain at crucial moments in the lead up to the vote. *Or* they could much more easily—and much more usefully—note the two or three key considerations weighing with the majority of voting members, expressed explicitly, as they might well be, in terms of beliefs and desires: "The Council *believes* that this action is necessary to preserve the amenity of recreational facilities in the area, which it *wishes* to safeguard ...."

This is not to say that deeper explanations of action are always irrelevant to normative assessment (see section 5). The point is rather that the cases in which they *are* required for assessing an action's merits are exceptional, especially where a compressed explanation facilitates accurate prediction (see below). At the limit, a formal "glass box" explanation of a network's decisions would provide so much detail as to obscure what really needs to be conveyed to the decision subject—who must, in the end, be leveled with as a fellow agent—namely, the semantic rationale of the decision, which allows the subject to gain an appreciation "of how and why things work the way they do and *what they mean*" (Leslie 2019, 40, emphasis added). In other words, brevity is important to justification insofar as it allows the subject to form a genuine *understanding* of why an action was taken.

Secondly, the behaviour-tracking feature of intentional explanation enables us to make *accurate predictions* about the agent's future behaviour. More precisely, it enables us to predict, at a level of accuracy significantly better than chance, what outputs a system will yield from specified inputs, including inputs which, from the system's point of view, may not have been previously encountered. But how does our being able to predict what a rational agent *will* do relate to the explanation of actions *already done*? What does prediction have to do with justification?

It is helpful to note that, in the sphere of practical reason, predictions function as a kind of ex ante explanation and explanations as a kind of ex post prediction. (More precisely, predicting what a system will do—i.e., its policy—is equivalent to the problem facing the agent who must, in light of its objectives, determine which is the most rational policy to adopt; hence it is equivalent to the problem of reinforcement learning: *f* (*objective*, *environment*) = *policy*. On the other hand, explaining what a system *has done* is equivalent to the problem facing the *observer* of an action who seeks to infer, on the basis of the policy ultimately adopted, the theory of mind most likely

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

to characterise the agent; hence it is equivalent to the problem of *inverse* reinforcement learning: f^{-1} (*policy*) = *objective*, *environment*.) To proclaim an action was justified, we would need to know why it was done, but knowing why it was done requires knowing that the particular set of factors adduced to explain the action was *in fact* operative; i.e., the set of factors from which we could reliably predict the same outcome if events were replayed. Predictions thus serve to verify explanations, and common intuitions reflect this. It is understood, for instance, that explanations under law "should permit an observer to determine the extent to which a particular input was determinative or influential on the output" (Doshi-Velez and Kortz 2017, 3). So, in citing beliefs and desires—items that *can* support reasonable predictions, assuming the agent in question is rational and truthful[5]—intentional explanations fulfill a basic requirement of justifying explanations.

At this point, however, one might wonder *how* a standard of accuracy could be settled upon for prediction; i.e., what does predicting "reliably," or "at a level of accuracy significantly better than chance," actually mean when the explanandum is a neural network? Given that—so far as we are concerned—explanations of such systems are intended for ultimate consumption by decision subjects, it makes sense 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. Of course, it may be impossible to determine quantitatively just how faithfully human reasons reflect human motivations: perhaps knowing the putative reasons why a human decision-maker chooses as they do rarely enables an observer to predict such choices retrospectively. But there is no basis for thinking that the "reasons" of a supervised ML system would be even less faithful guides to its behaviour than human reasons are to human behaviour. For one thing, such systems are not sophisticated enough to harbour deceptive motivations, so the assumption of good faith is, at least for them, unproblematic. Additionally, such systems—for all that they may be Type 2 systems—are unlikely to be more complex than human brains.

Notice that predictive power alone would be insufficient to account for action. In the right circumstances, one could merely simulate an agent's behaviour, predict with high confidence how it would behave when presented with particular inputs, and be no closer to understanding why it behaved in that way. The compression forced on an explanation by dint of its intentional structure allows the explanation to assume a more meaningful form.

Notice also that nothing I have so far said implies that intentional explanations are *ultimately* justifying. Clearly, for a decision to count as a *good* decision, its explanation needs to be more than just interpretable and verifiable; it also needs to be reasonable or advisable according to the standards of some domain of expertise or knowledge. Intentional explanations may be sufficient for justification only in the sense that they would enable someone pre-equipped with relevant domain knowledge—knowledge of the norms of competent chess-playing, or fair and reasonable criminal sentencing, or prudent commercial lending practice, or whatever—to assess a decision's merits. In this sense, "justifying explanations" are explanations that merely *purport* to justify an agent's action.

[5] See the next paragraph and section 4.5.

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

Together these two properties of an explanation, i.e., its interpretability (brevity) and verifiability (predictive accuracy), imply that the absence of a comprehensive account of action need have no bearing on an explanation's acceptability. Furthermore, that an explanation having such properties does not faithfully represent the subdoxastic states it is presumed to index need be no impediment to its acceptability either: there is nothing to prevent an intentional explanation being highly idealized, so long as it can reliably track the behaviour sought to be explained.

4.4 An intentional stance on machine learning?

It is one thing to say that the intentional stance pays off where humans are concerned. It is quite another to say that it pays off when dealing with ML systems. My aim in this subsection is to make more headway on the idea that we can adopt an intentional stance towards ML systems—i.e., to show in what sense it could pay dividends to do so. This is just to make the claim that ML systems are *rational* (much like Dennett's example-in-chief, the chess-playing computer programme). (Once again, the observations following relate to supervised learning systems, and neural networks in particular.)

But first, it may be wise to clear the air. I do not maintain that machines literally possess beliefs and desires, nor do I intend the use of such terms to be merely figurative or metaphorical. The idea is rather that by adopting the intentional stance towards a machine, something *like* what plays the role of a belief or desire in the interpretation of a person's behaviour could be seen to play that same role in the interpretation of the machine's behaviour. This is perhaps more aptly described as an *analogical* use, for it assumes that it makes sense to attribute the functional role equivalents—analogues—of beliefs and desires to any object for which the intentional strategy pays off. If the strategy does pay off, no further metaphysical assumptions need be made about what, *au fond*, beliefs are.[6] In like fashion, although Dennett often places terms like "belief" and "desire" in scare quotes when referring to the beliefs and desires of artifacts—suggesting he has a merely figurative or metaphorical sense in mind—strictly speaking, he considers a distinction between literal and metaphorical applications of intentional vocabulary to be "ill-motivated" (2009, 342, 343). This is consistent with his using the terms in what I have called an analogical manner (i.e., functionally). Obviously, the precise nature of a belief will differ from agent-kind to agent-kind, and indeed from agent to agent, but as Dennett notes, intentional systems theory allows us to set aside the "standard connotations" of mentalistic terms precisely "in the interests of exploiting their central features: their role in practical reasoning, and hence in the prediction of the behaviour of practical reasoners" (Dennett 2009, 339).

What could the neural network analogues of beliefs and desires be? If we adopt the intentional stance, the answer is obvious: whatever beliefs and desires it would make

[6] The beliefs we ascribe when adopting the intentional stance do not purport to exist *as such* somewhere deeper below the surface of behaviour, although Jerry Fodor's "industrial-strength Realism" does assume as much (see Fodor 1987; Dennett 1991). In a word, *that* the intentional stance pays off is more important to the issue of justification than *why* it pays off. Whether it works because belief-desire attributions correspond to real counterparts in the model, or (as Dennett would have it) merely to "patterns" in the model, or to schemata under some theory of instrumental explanation, the issue of concern for justification is whether the attribution itself makes the job of predicting a system's behaviour any easier.

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

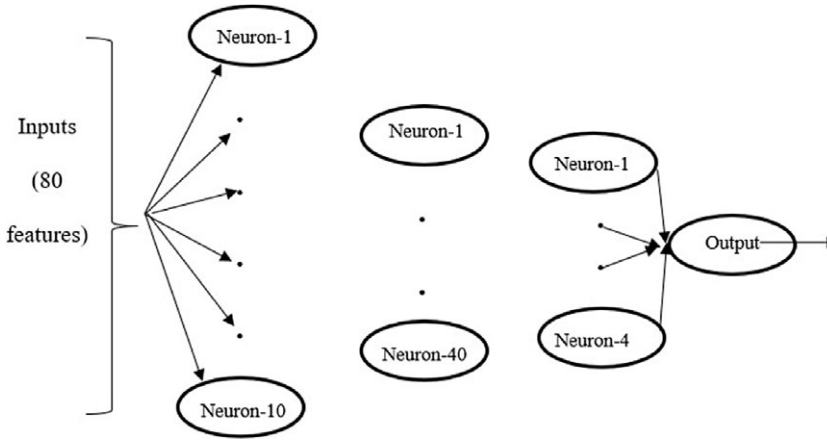

Figure 1. A three-layer neural network that receives 80 recidivism-relevant input variables to classify offenders as either posing recidivism risk or not. *Source:* Ozkan 2017, 49.

sense to attribute to a human agent performing the same task as the network. The more important question is whether the attribution of such states makes the job of interpreting the behaviour of an ML system any easier. I think the answer is "yes," but it is worthwhile noting how beliefs and desires may fare somewhat differently on this measure.

Take a neural network that assigns risk scores to prison inmates who are up for parole. It might compute over such variables as past offence type/s (on a scale of seriousness), current offence type, number of prior arrests, age at first arrest, sex, and age. Ozkan (2017) describes a network that has three hidden layers (each consisting of 10, 40, and 4 neurons each) whose first layer takes 80 such recidivism-relevant input variables (see figure 1). These inputs function in much the same way as a player's chess moves in a chess programme, in that they form part of the roster of the system's beliefs regarding the situation it is confronting. But there will be more in the network's roster than such input variables can account for, just as in the chess programme there were, in addition to beliefs about the position of the pieces, other beliefs that were *built in* to the system (rather than *fed in*), such as beliefs about which moves in the game are legal and which moves are optimal given particular configurations of the board. For a network trained on a sufficiently large set of previous offender statistics, the equivalent of its built-in beliefs will necessarily encompass all the generalisations it acquired during learning (and represents implicitly in patterns of synaptic weights). Very crudely, such generalisations might include the following:

```
IF offence seriousness > m AND age > b THEN high-risk
ELSE low-risk

IF EpisodeCount > n THEN high-risk
ELSE low-risk

IF behaved well THEN low risk
ELSE high risk
```

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

Unlike with chess programmes and expert systems, however, the main challenge for XAI lies precisely in tabulating that part of the roster of a network's beliefs that contains its learned generalisations (see section 4.5).[7] That the discovery of these generalisations would be a boon for any attempt at predicting how the network will behave is obvious (even if, as in the chess example earlier, such predictions are rarely foolproof). For instance, given just a few facts about an offender, the job of predicting what score a risk assessment tool will assign to them is easy if we assume that the above generalisations constitute the tool's entire belief system. If the category of offence seriousness ranges from 1 to 10, $m = 8$, and $b = 21$, we can easily predict that a well-behaved 18-year-old first-time offender who commits a category 7 offence will be classed as low risk. The example is rather simplistic, but it is not altogether fanciful. One risk assessment tool used by various state law enforcement agencies in the United States was found to be predictable with as few variables as the offender's age and number of previous convictions, despite explicitly utilizing 137 variables (Dressel and Farid 2018).[8] And to underscore the point that such beliefs are usually sufficient to assess justification, imagine if the tool correlated your Hispanic heritage, and sibling with a previous conviction for union picketing, with high criminality. These beliefs would be enough to mount a case against the legality of its decision to assign you a high-risk score.

Desires pose a different problem. When understood as agent goals, they can be quite simple (for a chess programme, "win the game," for a thermostat, "keep the temperature at 22 degrees Celsius," etc.). The desires (or goals) of supervised ML networks are to make very specific kinds of predictions. For a recidivism risk assessment tool, for example, the goal would be to assign a risk score to a prisoner, and the objective to correlate the system's various beliefs about the prisoner with their likelihood of reoffending if released. But these might be thought too simple to stand in for desires, and far from being too hard to formulate, they may be thought too easy.

It is true that the aim to "assign a risk score" is not a goal state of the kind we could relatively easily attribute to a chess-playing computer programme, or even a thermostat—a state, in other words, which the system *wants to be in* (in the sense that it seeks to maintain the state in the face of perturbation), achieved by means of a reinforcement learning algorithm (in the case of a chess programme) or simple feedback/PID controller (in the case of a thermostat). A recidivism tool simply reproduces the responses it is trained on. And while it could be said that it *wants* to reproduce such response patterns, this reduces to saying that the system's goal is "to do [what the system does]."[9] None of this, however, implies that the tool lacks goals in the relevant sense. Recall that the central premise of intentional systems

[7] It is also worth bearing in mind that such beliefs are unlikely to hold globally rather than merely locally (see section 4.5 on "local" explanation). In other words, because these beliefs are primarily intended to be useful insofar as they shed light on specific decisions made by a network, they may not hold more generally, and may, therefore, contradict other (global) beliefs of the system.

[8] As it happens, the tool Dressel and Farid discuss does not (as far as we know) employ a neural network to generate risk scores, but this has little bearing on the point I am making here, which is simply that a belief system need not be especially rich to confer genuine predictive insight into the behaviour of the agent whose beliefs it describes.

[9] On the other hand, we could also say that if the network encodes a strong belief that the current situation is of category *C*, because of some input *v*, but other input *w* tries and fails to override this belief,

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

theory is that if by treating a system as having beliefs and desires one is enabled to predict its behaviour more efficiently than by attending to features of its constitution and design, then one can simply say that it *has* those beliefs and desires. And the fact remains that only in light of *both* its beliefs about recidivism *and* its broader functional rationale can we make predictions about what score the tool will assign in a given case. For, in addition to the system's roster of beliefs, it is the system's broader purpose or function of assigning risk scores to prisoners presenting with particular criminal histories that structures, however trivially or innocuously, our expectations regarding the system's behaviour. The fact that I believe that attributes *p*, *q*, and *r* correlate strongly with recidivism does not tell you what I will do next, unless you also know that my job involves sitting on a parole board and that my aim is to determine, as fairly and accurately as possible, whether prisoners coming up for parole should be released. That my goal, trivially, is "to do [my job]," does not mean that my goal does no extra work on top of what I happen to believe about how best to do my job. Indeed, though it is natural to do so, to assume that my goal as a parole officer is to give effect to my beliefs about recidivism by forming accurate assessments of risk is not without its hazards. I may be venal or have ulterior motives besides. I may, for instance, have an undisclosed association with a prisoner from whose clement treatment I stand to benefit—despite what my beliefs about recidivism imply about their risk of reoffending. If so, my parole determinations will remain obscure until my true motivations are exposed. ML systems do not have such ulterior motives, of course, so in knowing a system's beliefs about recidivism it is always safe to assume that its goal will be to give effect to those beliefs. But the assumption's safety in the circumstances does not negate the distinction between a system's beliefs and desires. The nub of the worry, then, is not so much that supervised ML systems lack the sorts of aims that make adopting the intentional stance worthwhile, as that their desire states are going to be somewhat trivial when compared with their vastly richer belief states. But this much can be conceded.

4.5 Examples of interpretable explanations in XAI

XAI is a heterogeneous movement, but a significant focus of research within it centres on the production of post hoc explanations (Rudin 2019; Leslie 2019; Guidotti 2018; Adadi and Berrada 2018; Selbst and Barocas 2018; Lipton 2017). These, in essence, are simplified models of an underlying model explanandum ("models of a model"), as a consequence of which they tend to be partial, schematic, and idealized: they do not necessarily aspire to provide models that mimic the actual logic of the systems being explained (Rudin 2019; Leslie 2019; Lipton 2017). An important strand of this enterprise, in turn, can be understood as attempting to formulate a network's built-in beliefs accurately, so that reliable predictions can be made about what it will recommend given a set of inputs (i.e., its fed-in beliefs). We could cite "explanations by example" as a case in point, in which a network's classification of a tumour as malignant might be explained by reference to the similarities the tumour exhibits with others the network has been trained to recognise as malignant. As Lipton

the network *wants* to assign category *C* in a sense of "want" more approximate to the chess programme's or thermostat's.

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

(2017, 6) observes: "This sort of explanation ... has precedent in how humans sometimes justify actions by analogy. For example, doctors often refer to case studies to support a planned treatment protocol." But equally striking is the intentional structure of the explanation. The network's classification is, in effect, explained in terms of its *belief* that the tumour looks like other tumours it *believes* to be malignant.

A subset of post hoc techniques is described as "local" or "ex post," meaning that instead of treating the system *as a whole* as an explanandum, these techniques treat a system's *specific* decisions or outputs as the explanandum (Leslie 2019; Selbst and Barocas 2018). The thought underlying this approach is that despite what may be the very real complexity of a decision function, it can, nonetheless, be interpreted through analysis of specific data points or regions within its larger feature space (Leslie 2019). Local interpretability methods, therefore, seek to establish a feature's importance to a decision, and they do this by "iteratively varying the value of that feature while holding the value of other features constant" (Selbst and Barocas 2018, 1114). Sensitivity analysis, for example, tries to gauge which of the various features comprising an input vector $\mathbf{x_i}$ has the greatest bearing on an output variable. For our dog and cat classifier, this would amount to revealing which of the set of input-output pairs ($\mathbf{x_i}$, y_i) has the strongest association. While the question is answered *ex post* for specific output variables (i.e., only in the wake of specific classifications), the technique does have the potential to reveal a network's learned generalisations by revealing that the network encodes strong beliefs that this or that feature of an image is strongly suggestive of a cat and not a dog (for example). Much the same could be said for the related technique of saliency mapping, except that the network's beliefs are represented visually in a heat or pixel attribution map. Perhaps the most high-profile local interpretability method is LIME (Local Interpretable Model-Agnostic Explanation). Ribeiro et al. (2016) were able to show that a deep learning model trained on images of wolves and huskies used the presence or absence of snow in an image to distinguish between them (a classic case of a model overfitting its training data—reflecting the fact that wolf images in the training set generally had snow in the background, while husky images did not). This is a very pure case of intentional explanation ("I believe that snow in an image indicates a wolf; image X had snow in it, so I concluded it must be a wolf"). It also starkly exhibits the essential ingredients of a (putatively) justifying explanation. Because snow is not a property of the class WOLF, the cited reason for the classification is spurious, and can be dismissed as unjustified. But crucially, the explanation gives us everything we need to determine this.

Each of these techniques exemplifies the kind of explanation that I have argued is often sufficient for assessing justification. More generally, the explanations mirror the format of interpersonal explanations, including those promoting accountability under law, which are "about answering how certain factors were used to come to the outcome in a specific situation ... rather than an explanation of the system's behavior overall" (Doshi-Velez and Kortz 2017, 7). Still, this is not to say that these explanations are always *ideal* exemplars of the strategy I have defended. For instance, the local methods so far devised will likely fall short where a feature space incorporates many features that contribute equivalently to an outcome (Selbst and Barocas 2018). Listing any fewer than the correct number of features in such a case may not allow us to generate reliable predictions, but listing the correct number could defeat

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

the purpose of striving for an interpretable explanation in the first place. In another vein, a local explanation would be misleading if it tracked an excluded feature instead of a proxy that actually determined a model's prediction. Thus, a post hoc explanation of a recidivism model in which criminal history and age are correlated with race, but in which race itself does not feature, would be in some measure misleading if the explanation cited race as the reason for a high-risk score. (However, so long as it predicted risk scores accurately by citing race, it would be a valuable explanation nonetheless: it would be latching onto the fact that criminal history and age are, for a variety of sociological reasons—including discrimination—correlated with race in most datasets.) LIME too, has received its fair share of criticism, mostly revolving around such issues as how best to define the region of the input space to which its explanations apply and the distorting effects of very small perturbations on the underlying model. But while worries about local and other post hoc methods generating misleading explanations cannot be lightly brushed aside, it is important to be clear about why they risk being misleading. It is not because they are idealized, or (pace Rudin 2019, 3) lack "perfect fidelity with respect to the original model."

I conclude this section by acknowledging that an intentional approach to interpretability is not the only one available. It can be set against another approach in which interpretable accounts of ML systems may be given that do not resort to agentive vocabulary at all. The most obvious examples here would be the explanations of fathomable and other intelligible systems—systems, in other words, utilising regression techniques, decision trees, rule lists, and the like. It may be easy to explain how a linear regression model has optimized its parameters, and thus to explain why the model has generated a particular outcome in a particular case, and yet it would be a mistake to assume that such ease of exposition must be attributable to the use of agentive vocabulary. But the same can actually be said of Type 2 systems. Many of the pragmatic post hoc techniques that were advocated in the 1980s for connectionist systems (see section 2.2) could be understood as generating interpretable explanations that were not *also* intentional. These explanations would, presumably no less than intentional explanations, suffice to enable an assessment of an automated decision's justification. On my account, what is crucial for explanations of ML decisions is brevity and predictive accuracy, so if some *non*-intentional explanations of ML decisions happen to offer these same features, so much the better for them (i.e., they too may be sufficient to justify action).

5. When deeper explanations are needed

5.1 Outline

There certainly are occasions when design-level explanations of automated decisions will be required before a proper assessment of their justification is possible. The most obvious would be where an intentional explanation devised in XAI was not yet up to scratch, for example, because it did not facilitate reliable prediction. But there are two important cases worth mentioning apart from this where deeper (design-level) explanations of automated decisions will be required before their justification can be assessed: the first has to do with the degree of sophistication of the system in question, the second with gaps in our knowledge of the context in which a system is deployed.

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

5.2 Simplicity

Although decision systems might unproblematically qualify as functional kinds, and be multiply realizable in principle, it may be that some putative ML decision systems simply cannot count as genuine realizations of such systems, there being nothing they would be able to generate at the level of practical deliberation that adequately resembles a "decision." Perhaps an elementary rule of thumb might then be that, by and large, the simpler that a decision system is—i.e., the less proximate to a "decision" its deliberations amount to—*the deeper the explanation we should expect from it.*[10] This accords with prevailing practices around linear and simple logic-based decision systems. Their operations are fathomable precisely because they are simple, but perhaps it is because they are simple that we have demanded the exhaustive explanations from them that they readily, as it happens, provide. Putting the point slightly differently: these systems qualify as rational by the lights of intentional systems theory, and so *can* have their decisions adequately assessed via intentional explanation; but perhaps it is this very simplicity that renders their intentional explanation adequate—they are simple enough for their intentional explanation to offer a deep account of their operations.

5.3 Knowledge gaps

In some cases, we may know what reasons a system has for deciding something but be none the wiser about what to make of them. This will be the case where an automated decision system recommends a course of action on the basis of a correlation that appears contrary to human intuition and that cannot further be illuminated by existing knowledge (Selbst and Barocas 2018).

Often a baffling association can be explained away without too much trouble. A system might, for instance, detect a statistically significant correlation between asthma and recovery from pneumonia. The true cause is unlikely to be that asthmatics are more robust to infection, but rather that they are targeted for more aggressive treatment upon clinical presentation with pneumonia, which then enhances their odds of recovery (Caruana et al. 2015). The full explanation of the system's mechanism of discovery would encompass information not just about the existence of the association, but about what common third factor likely accounts for it. Such an explanation would then furnish all the resources needed to determine whether a recommendation of the system was justified.

The problem, however, is that not all counterintuitive associations can be explained away by further evidence that some third factor mediates between correlated variables. Sometimes a full explanation—so far as existing knowledge can provide it—stops at the discovery of the correlation. In these cases, even a fairly rich explanation of the system may not provide a reviewer with enough information to assess the value of the explanation, and thus to assess whether any decision of the system was justified. Yet it would be rash to dismiss the system's findings tout court just because they did not comport with human intuition.

[10] Obviously, this rule will not apply when the stakes are low: to clinch some matters, we might be happy to toss a coin, and would never insist on being able to obtain a full reckoning of the coin's trajectory before doing so.

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

5.4 Malfunctions and alternative explanatory goals

A different consideration entirely is the potential for system malfunctions, which may require intensive understanding of a system before they can be remediated. More generally, design-level explanations are sought in the interests of control (i.e., to *prevent* malfunctions) and improving performance, as well as for general understanding (Adadi and Berrada 2018).

6. Conclusion

I have argued that in accounting for the decisions of automated systems, it will not always be necessary to have recourse to deep and exhaustive accounts of the system's operations—for example, of the kind to which we became accustomed when much simpler ML and logic-based systems were the norm. These simpler systems divulged their inner processing logic to a degree that it is not reasonable, or even necessarily desirable, to expect from today's neural networks. So long as we are concerned with the evaluation of decisions, the formal prerequisites of explanation are in many cases satisfied by the quotidian form of explanation that practical reasoning assumes. That human practical reasoning takes this form is instructive for the kinds of explanations we can reasonably demand of ML systems, because such systems function in loco hominum, and indeed have parity with humans not only in respect of how they are situated, as decision agents, but also in respect of the epistemic status they afford, as Type 2 systems. When the necessity for deeper and more comprehensive explanations of automated decisions is urgent, as in some cases it may be, we should naturally expect them, in whatever form is considered practicable by the standards of XAI. But where no such necessity arises, a satisficing explanation of an automated decision ought to suffice for assessing its credentials.

References

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black Box: A Survey on Explainable Artificial Intelligence." *IEEE Access* 6:52138–2160.

Boden, Margaret. 1990. *The Philosophy of Artificial Intelligence*. New York: Oxford University Press.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." *Proceedings of the 21st ACM International Conference on Knowledge Discovery and Data Mining,* 1721–730.

Clark, Andy. 1990. "Connectionism, Competence, and Explanation." *British Journal for the Philosophy of Science* 41:195–222.

Dennett, Daniel C. 1971. "Intentional Systems." *Journal of Philosophy* 68 (4):87–106.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, Daniel C. 1991. "Real Patterns." *Journal of Philosophy* 87:27–51.

Dennett, Daniel C. 2009. "Intentional Systems Theory." In *The Oxford Handbook of Philosophy of Mind*, ed. A. Beckermann, B.P. McLaughlin, and S. Walter, 339–50. New York: Oxford University Press.

Doshi-Velez, Finale, and Mason Kortz. 2017. "Accountability of AI Under the Law: The Role of Explanation." Version 1. https://arxiv.org/pdf/1711.01134v1.pdf

Dressel, Julia, and Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4:1–5.

Fodor, Jerry A. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51 (5):Art 93, 1-42.

Leslie, David. 2019. *Understanding Artificial Intelligence Ethics and Safety*. London: Alan Turing Institute.

https://doi.org/10.1017/psa.2021.13 Published online by Cambridge University Press

Lipton, Zachary C. 2017. "The Mythos of Model Interpretability." ICML Workshop on Human Interpretability in Machine Learning. https://arxiv.org/pdf/1606.03490.pdf

Marr, David. 1977. "Artificial Intelligence: A Personal View." *Artificial Intelligence* 9:37–48.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Ozkan, Turgut. 2017. "Predicting Recidivism through Machine Learning." PhD diss. University of Texas, Dallas.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining*, 1135–44.

Rosch, Eleanor. 1978. "Principles of Categorization." In *Cognition and Categorization*, ed. E. Rosch and B.B. Lloyd, 27–48. Hillsdale: Lawrence Erlbaum Associates.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1:206–15.

Selbst, Andrew D., and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines." *Fordham Law Review* 87:1085–139.

Cite this article: Zerilli, John. 2022. "Explaining Machine Learning Decisions." *Philosophy of Science* 89 (1):1–19. https://doi.org/10.1017/psa.2021.13