Title: The Limits of Value Transparency in Machine Learning

Abstract: Transparency has been proposed as a way of handling value-ladenness in machine learning (ML). This paper highlights limits to this strategy. I distinguish three kinds of transparency: epistemic transparency, retrospective value transparency, and prospective value transparency. This corresponds to different approaches to transparency in ML, including so-called 'Explainable AI' and governance based on disclosing information about the design process. I discuss three sources of value-ladenness in ML—problem formulation, inductive risk, and specification gaming—and argue that retrospective value transparency is only well-suited for dealing with the first, while the third raises serious challenges even for prospective value transparency.

Contact information: Rune Nyrup, Leverhulme Centre for the Future of Intelligence, University of Cambridge, rn330@cam.ac.uk

Acknowledgements: Many thanks to William Peden, Ali Boyle and Milena Ivanova for providing very helpful comments on a draft of this paper at short notice. In addition to the PSA2020/21 symposium "Philosophy of Science Meets AI Ethics", previous versions were presented at the "Philosophie und Wissenschaftsreflexion Colloquium", Lebniz Universität Hannover, "Virtual Workshop on the Philosophy of Medical AI", University of Tübingen, and the 2018 BSPS in Oxford. I'm grateful to the audiences at these events for their valuable feedback. This research was funded in whole, or in part, by the Wellcome Trust [Grant number 213660/Z/18/Z] and the Leverhulme Trust, through the Leverhulme Centre for the

Future of Intelligence.

The Limits of Value Transparency in Machine Learning

Abstract: Transparency has been proposed as a way of handling value-ladenness in machine learning (ML). This paper highlights limits to this strategy. I distinguish three kinds of transparency: epistemic transparency, retrospective value transparency, and prospective value transparency. This corresponds to different approaches to transparency in ML, including so-called ‘Explainable AI’ and governance based on disclosing information about the design process. I discuss three sources of value-ladenness in ML—problem formulation, inductive risk, and specification gaming—and argue that retrospective value transparency is only well-suited for dealing with the first, while the third raises serious challenges even for prospective value transparency.

1. Introduction

Computer ethicists have long argued that computer systems are value-laden (Friedman and Nissenbaum 1996; Kraemer et al. 2011). Machine learning (ML) is no exception, as highlighted by investigative journalists and academic studies alike (Angwin et al. 2016; Obermeyer et al. 2019). Recently, philosophers of science have begun to address this issue, drawing on parallels with the values in science literature (Biddle 2020; Johnson forthcoming). This paper further explores these parallels, focusing on a strategy that has been proposed for managing value-ladenness in both domains: *transparency*.

Scientific results are often the result of complex chains of justification. This can occlude significant value-laden decisions to non-experts who rely on these results. Philosophers have argued that to mitigate the impacts of value-ladenness scientists should make transparent either the uncertainties (Betz 2017) or the values involved in such decisions (Douglas 2009; Elliott 2017; 2020). Though not uncontested (Schroeder 2021; Nguyen 2021), proponents of transparency argue that it helps secure democratic legitimacy and autonomy for non-experts (Elliott 2010).

Similar concerns drive calls for transparency in ML. As with scientific results, the complexity of modern ML systems, and their ability to discover novel correlations, make them epistemically opaque (Burrell 2016; Sullivan 2022; Zednik 2021). A growing field of technical research, known as ‘interpretable’ or ‘explainable’ AI (XAI), seeks to improve the epistemic transparency of ML systems (Zednik 2021; Zerilli 2022). However, many law and technology scholars argue that this focus is misplaced, proposing instead governance frameworks based on making transparent the goals and values that ML systems promote (Selbst and Barocas 2018; Kroll 2018).

Drawing on the values in science literature, this paper evaluates these different approaches to transparency in ML. I start by outlining my preferred construal of value-ladenness and distinguish three types of transparency: epistemic transparency, retrospective value transparency and prospective value transparency. Next, I apply this distinction to transparency in ML. I highlight limitations of XAI as a means to promoting epistemic transparency. Finally, I consider three sources of value-ladenness in ML: problem formulation, inductive risk, and specification gaming. I argue that retrospective value transparency is only well-suited for dealing with the first, while the third raises serious challenges even for prospective value transparency.

2. Value-Ladenness

Scientific inquiry involves choices that could reasonably have been made differently. Scientists make decisions about how to conceptualise and frame research questions, what type of models to develop, which hypotheses to test, how to collect data and design experiments, and what kinds of evidence suffice to accept or reject hypotheses. There are often several available options, each with advantages and drawbacks, but no decisive reason to prefer one over the other. In these cases, the decisions scientists make are *contingent* (Brown 2020, 57-86): someone in the same epistemic situation—with the same evidence, background information, methodologies, available equipment, and other resources for inquiry—could reasonably have chosen differently.

A contingent decision is value-laden when the reasonable alternative options differ in their potential impacts on the things we value and care about. One much-discussed form of value-ladenness is inductive risk, i.e., the risk of inferential error. Since different types of errors—most saliently false positives and false negatives—typically differ in their potential impacts, methodological decisions which affect the balance of inductive risk are often value-laden (Douglas 2009). Another example concerns the types of evidence different methodologies provide. For example, randomised controlled trials tend to be good for assessing whether a given intervention produced some population-level outcome but are relatively uninformative about distributive consequences. Thus, if most researchers pursue randomised controlled trials, it becomes difficult for policymakers to find evidence relevant to policies aimed at distributive values, such as equality or priority for the worst-off (Khosrowi 2019).

There are two things worth noticing about this definition of value-ladenness. First, it makes *decisions* the primary locus of value-ladenness. However, *products* of inquiry, such as theories, models, or algorithms, can be called value-laden in a derived sense if they are shaped by and mediate the potential impacts of value-laden decisions. That is, if some contingent decision had been made differently, the product would been different in a way that changes its potential impacts on the things we value.

Second, decisions are defined as value-laden regardless of whether value judgements played any role in reaching or justifying them. As Ward (2021) argues, philosophers have discussed (and often conflated) four distinct kinds of value-ladenness: (i) decisions consciously motivated by values, (ii) decisions justified by values, (iii) decisions causally influenced by values (e.g., through implicit biases or institutional structures), and (iv) decisions that impact the things we value. While there are often complex interactions between these, my use of the term in this paper is restricted to (iv).

Few philosophers would deny that many scientific decisions are value-laden in this sense (though spelling out exactly how is often not trivial; Ward [2021, 57]). Rather, framed this way, the main disagreements concern what can and should be done to manage value-ladenness. I will focus on issues relating to transparency. These arise when scientific products have been shaped by value-laden decisions in ways that are unclear to those relying on them. Take randomised controlled trials again: it would be problematic if contingent decisions about what kinds of evidence to prioritise systematically inhibited policymakers from pursuing otherwise legitimate values, especially if this happened without the awareness of policymakers or the public. More generally, if non-experts are unaware of the value-ladenness of the scientific products they rely on, it risks undermining their ability to determine which values are prioritised in their own decision making (Betz 2017; Elliott 2010; Schroeder 2021). Concerns of this kind motivate calls for transparency.

3. Epistemic Transparency and Value Transparency

The concept of transparency is complex (Biddle 2020; Elliott 2020), but generally involves providing relevant kinds of information to non-experts. I take this to involve at least two things:

(a) *Openness*: communicating (or making accessible) the relevant information to the right audience.

(b) *Comprehensibility:* ensuring the audience can understand and use the information.

These in turn entail a further precondition:

(c) *Explicitness*: ensuring the communicator is aware of and able to articulate the information in the right way, i.e., so that it achieves (a) and (b).

Different kinds of transparency can be distinguished based on the type of information involved. Philosophers have promoted two general kinds of transparency in response to value-ladenness: *epistemic transparency* and *value transparency*.

Epistemic transparency focuses on information about the uncertainties or justifications involved in contingent decisions. For instance, Betz (2017) proposes *full uncertainty disclosure* as a strategy for managing inductive risk. Briefly, he argues that scientists should only endorse claims that are beyond reasonable doubt. Instead of asserting uncertain conclusions, these should be reframed as “hedged hypotheses”, where the type and degree of uncertainty are made explicit. When faced with contingent choices, rather than picking one option, scientists should analyse as many as possible, report the consequences of each option, and specify any remaining uncertainties.

The motivation for epistemic transparency is to leave as many value-laden decisions as possible to those who rely on scientific products. Ideally, scientists should merely lay out the options and explain uncertainties, leaving it to non-experts to decide what risks they are willing to take, given the values they want to prioritise. However, this strategy faces limitations in relation to each of (a)-(c). First, due to the number of contingent choices scientists face, it easily becomes intractable to analyse and communicate all but a small fraction. Second, the technical vocabulary necessary to accurately specify the relevant uncertainties may not be comprehensible to non-experts. (Betz for instance mentions "imprecise probabilities, fuzzy logic, or degrees of possibility" [2017, 104]). Third, parts of scientists' knowledge are arguably tacit, situated in know-how, habits, or social structures. Scientists may not themselves be able to make these fully explicit without distortion (Nguyen 2021). To be sure, Betz emphasises that his view is only meant as an ideal norm. But this leaves open how to manage value-ladenness under non-ideal circumstances.

Value transparency focuses on informing non-experts about the values involved in value-laden decisions or products. Rather than leaving it to non-experts to make all value-laden decisions for themselves, this strategy permits scientists to make such decisions, provided they inform the non-experts about the relevant values. Elliott (2020, 3) outlines several motivations for this kind of transparency. First, it would warn non-experts that relying on a given scientific product risks promoting values they disagree with. Second, it may allow them to reanalyse or reframe results in ways that better accord with their priorities. Third, it can enable non-experts to influence what decisions are made in the first place (Douglas 2009, 153). Finally, it can play an important role in supporting other strategies for managing value-ladenness, such as promoting engagement with different publics who might be impacted (Elliott 2017, 137-162).

The above characterisation is meant to be consistent with all four of Ward's senses of value-ladenness. Nonetheless, it is useful to distinguish two sub-types: *retrospective value transparency*, i.e., what values motivated and influenced a given decision (Ward's first and third senses), and *prospective value transparency*, i.e., what are the potential impacts on things we value of relying on a given decision/product and, relatedly, what values would justify doing so (Ward's second and fourth senses).

Some discussions of value transparency articulate it in mainly retrospective terms (e.g., Douglas 2009, 153; Elliott 2020, 3; Schroeder 2021, 550). However, the two can come apart. In many cases, what is directly relevant to non-experts is what values they risk promoting prospectively. Knowing what motivated or influenced decisions (e.g., through conflict-of-interest statements) at most provides an indirect way to gauge this. Prospective value transparency does not require detailed explanations of *how* decisions produce their potential impacts or *why* certain values would justify decisions. Rather, the point is to explain the conclusions of such analyses to non-experts, such that they can understand *what* the potential implications are for the values they care about.

This is not to say that prospective value transparency is without its challenges. Explicitness can be especially hard to achieve, as it requires the communicator to predict the likely consequences of scientists' decisions. Moreover, values can be difficult to express precisely. If scientists and non-experts have different conceptions of core value concepts (e.g., justice, freedom, health), miscommunication could easily ensue.

Ultimately, none of the three types of transparency distinguished in this section is a panacea. Rather, they are more plausibly seen as complimentary strategies that can be used within broader efforts to manage value-ladenness. Thus, as Elliott (2020) argues, evaluating the merits of transparency-based strategies will involve more detailed questions concerning the benefits of different types of transparency. In the rest of this paper, I will look at some such questions for the case of ML.

4. Explainable AI as Epistemic Transparency

A prominent objection to using ML in high-stakes decision making, the so-called “black box problem”, concerns epistemic opacity.[1] In addition to intentional secrecy by the organisations who control algorithms and lacking technical understanding among users, many have also argued that there are features specific to advanced ML systems that make them epistemically opaque.

ML systems are defined as computer programs capable of using data to improve their performance on some task (Mitchell 1997, 2). This is typically approached as an optimization problem. For example, many applications of supervised ML are based on search algorithms (e.g., gradient descent) which iteratively adjust the free parameters of a model to optimize its performance on a given dataset according to some performance measure.

Two things distinguish advanced forms of ML, as far as transparency is concerned, from this basic picture. First, the complexity of models. This is not just size—though some

[1] For earlier commentary, which partly informs my discussion, see Burrell (2016), Erasmus, Brunet, and Fisher (2021), Selbst and Barocas (2018), Sullivan (2022), Zednik (2021), Zerilli (2022).

ML models are huge, sporting millions or even billions of parameters—but also the highly non-linear relations that they are able to represent. When models involve many parameters that interact non-linearly, it becomes difficult to disentangle and make sense of the dependencies between inputs and outputs.

Second, and partly because of this expressive power, ML systems are often able to discover correlations in datasets that go beyond existing human knowledge and understanding. While the power of advanced ML systems to a large extent rests on their ability to find and exploit such correlations, it can also make it difficult to explain their performance (Sullivan 2022; Zednik 2021). Specifically, it inhibits our ability to explain what features of the target domain the model is tracking (because we lack knowledge of what its features are) and why tracking these features results in the observed model performance (because we lack understanding of the dependencies between them).

A growing body of technical research, called 'explainable AI' (XAI), seeks to develop tools for mitigating these challenges. These include methods for generating partial or idealized representations of ML models, as well as more localised representations of how different inputs relate to a given prediction. The latter are particularly relevant here. Many of these are based on sensitivity analyses: varying one or more input variables to estimate which inputs made the largest difference to a given prediction. These estimates can be presented as heatmaps for visual data, rankings of input features, or counterfactuals.[2] Another popular

[2] Zednik (2021), Zednik and Boelsen (forthcoming), and Zerilli (2022) discuss examples of these techniques.

approach is to provide uncertainty estimates for predictions, e.g., by approximating Bayesian posteriors (Gal and Ghahramani 2016) or using ensemble techniques to estimate (non-Bayesian) confidence intervals (Lakshminarayanan et al. 2017).

While of necessity brief, this overview suffices to illustrate the point I want to make here, namely that most current XAI tools aim at *epistemic transparency*: they seek to illuminate the justifications and uncertainties that underpin the predictions of ML systems. In fact, many XAI techniques aim at something close to what Betz recommends, by testing the sensitivity of decisions and explicitly reporting uncertainties. Moreover, although different motivations are cited in the XAI literature for using these techniques, the most important one is arguably to enable those relying on ML systems to assess whether their decisions are *justified*: that is, to enable people to assess whether decisions are made on a basis they consider normatively acceptable (Selbst and Barocas 2018, 1122–26; Zerilli 2022). Again, this resembles Betz's motivation for epistemic transparency, namely to enable non-experts to make the relevant value judgements for themselves.

It should therefore not come as a surprise that these techniques, considered as ways of managing value-ladenness, suffer from many of the same limitations as epistemic transparency in science. First, the sensitivity analyses underlying, e.g., heatmaps and counterfactuals only analyse a small subset of the potential alternatives and may misleadingly suggest that the same factors would also be important in other, similar cases (Selbst and Barocas 2018, 1113–15). Second, not all audiences will have the expertise to comprehend probabilistic uncertainty representations or technical definitions of counterfactuals. Finally, information about the features that influenced a given output will only allow the audience to evaluate its justification if paired with additional knowledge of how the features that input and output variables represent depend on each other, as well as how different kinds of predictions might impact things we value (Selbst and Barocas 2018, 1126–29). But, as mentioned, it is exactly the ability to discover correlations beyond our existing knowledge and understanding that underpins the reliability *and* opacity of many ML systems.

To be clear, this is not to say that XAI tools cannot be helpful for managing value-ladenness in ML. Rather, my point is that they cannot be the whole of the story.

5. The Limits of Value Transparency

Because of these limitations, some law and technology scholars are sceptical of XAI as a governance tool for ML. Instead, they advocate frameworks based either on providing information about value-laden decisions in the design of ML systems or assessments of their impacts.

For instance, Selbst and Barocas (2018, 1130) argue that designers should document information which can help others understand "(1) the values and constraints that shape the conceptualisation of the problem, (2) how these values and constraints inform the design of machine learning models and are ultimately reflected in them, and (3) how the outputs of models inform final decisions". In other words, this strategy relies on value transparency rather than epistemic transparency. More specifically, as they highlight values that motivate design choices, it is a version of retrospective value transparency.

Others have proposed governance schemes that include elements of prospective value transparency. For instance, Kroll (2018, 1) rejects the idea that ML systems are uniquely mysterious or unaccountable, arguing instead: “Software systems are designed to interact with the world in a controlled way and built or operated for a specific purpose, subject to choices and assumptions. … Technologies can always be understood at a higher level, intensionally in terms of their designs and operational goals and extensionally in terms of their inputs, outputs and outcomes”. For Kroll, assessing and monitoring the potential impacts of ML systems against well-defined criteria provides exactly the kind of transparency we should demand of any technology.

In the following, I discuss the prospects of both types of value transparency for managing three sources of value-ladenness in ML systems: problem formulation, inductive risk and specification gaming.

5.1 Problem Formulation

An important type of contingent decision in designing an ML system concerns what Passi and Barocas (2019, 39) call *problem formulation*, that is, “turning amorphous goals into well-specified problems” that can be solved using ML. While goals and problems expressed in natural language can often be vague or ambiguous, leaving many implicitly understood things unsaid, ML problems have to be fully and explicitly defined. This requires several decisions. For example: what type of ML problem should the goals be modelled as (e.g., regression problem, clustering problem, matching problem)? What target function should the system aim to capture? How are target variables operationalised as measurable data points? How should model performance be defined and measured? What cost function should optimisation algorithms try to minimise?

Most such decisions are value-laden. Different reasonable implementations of a given informal problem formulation will prioritise different desiderata and constraints. To use Passi and Barocas' example, to design a system for identifying good job applicants, designers will need to find a way of specifying what "good" means. Some aspects of being a good employee (e.g., being personable and good at teamwork) are difficult to define and measure directly. Designers could ask managers to hand-label CVs with their judgements of teamwork ability, but those judgements are rarely completely reliable and could easily reproduce social biases. Instead, designers might rely on readily measurable things like sales figures, even if this only captures one aspect of what the company is looking for. These decisions are clearly value-laden: they will impact whether the ML system is more likely to make reliable predictions but of a variable that is less faithful to the original problem, or vice versa.

Retrospective value transparency seems promising for managing this type of value-ladenness. Even if designers are not directly aware of how their decisions are value-laden, it is relatively close to the surface. Getting them to explain what their goals are for the system, why they translated these goals into a particular problem specification, and what the pros and cons of alternative decisions were, should provide good evidence of what the operative values are. While it may still take some analytical work to make these values explicit, in many cases—such as deciding to operationalise 'good employee' in terms of sales figures—it will be relatively straightforward.

5.2 Inductive Risk

Inductive risk is a perennial feature of ML (and indeed any algorithm; Kraemer et al. 2011). All real-life applications will involve some risk of error, and different design choices will impact what kinds of errors are more likely to occur and who they are more likely to affect.

For instance, certain variables might be more likely to encode pre-existing biases, as Obermeyer et al. (2019) show for a system deployed to prioritise patients with higher needs for various public health programmes. Here, the designers used healthcare spending as a proxy for healthcare needs, perhaps because it was a plentiful and readily available dataset. This turned out to systematically disadvantage Black patients, as they tend to have less money spent on their healthcare compared to other patients with the same disease burden. Using more direct measures of healthcare needs, such as disease burden, might mitigate this (although there could still be disparities in how they are reported). However, if this data is sparser it could also lead to higher overall error rates.

Inductive risk can to some extent be addressed through retrospective value transparency. For instance, designers can assign differential weights to false positives and false negatives in performance measures and objective functions. Similarly, ML researchers have proposed various statistical measures of ‘fairness’ which can be used to detect and mitigate disparities, say, in false positive rates between sub-populations (see Biddle 2020 for discussion). However, these techniques treat labels in the data set as ground truth; if the data itself is biased, this will not be reflected in performance measures (Passi and Barocas 2019, 40). Detecting this kind of bias thus requires an independent dataset, ideally one reflecting the real-world impacts directly. This is for instance what Angwin et al. (2016) tried to do when they looked at whether defendants categorised as high or low risk of recidivism by the COMPAS system in fact went on to commit new crime.

Fully managing inductive risk and bias thus requires us to look beyond the choices of designers themselves, and evaluate the impacts of ML systems directly. In other words, it also requires prospective value transparency. To be sure, once a given value-laden impact has been detected, it will often be possible to backtrack and attribute this to one or more decisions within the design process. But if those impacts were not predictable at the time, merely requiring transparency about designers' choices and motivations are unlikely to reveal them.

5.3 Unintended Solutions and Specification Gaming

This challenge becomes even more pronounced in the final source of value-ladenness I want to discuss. As mentioned, the power of advanced forms of ML lies in their ability to discover and exploit surprising correlations. Such surprises are often exactly what designers are looking for: they deploy ML because they lack an explicitly programmable procedure for solving a problem. Yet the ability to surprise simultaneously gives rise to the phenomenon known as *specification gaming* (Krakovna et al. 2020): when an ML system discovers a "solution" to the problem as it was formally specified which is however undesirable given the designers' intended goals.

ML researchers have demonstrated many cases of specification gaming.[3] Some involve ML agents designed to play computer games. For instance, one such agent learnt to exploit bugs in the software which enabled it to crash the game just before losing a match. But examples of specification gaming have also been documented in real-world applications. To take a simple example, a system designed to classify skin lesions learned that malignant lesions in the training data were more likely to have been photographed next to a ruler (Patel 2017). Obviously, this correlation would not have been a robust basis for making predictions.

The illustrative examples of specification gaming involve outright failures (games crashing, unreliable prediction rules). However, it could equally lead to subtler forms of value-ladenness, e.g., by finding a solution which implicitly prioritises one plausible interpretation of the motivating goals.

Specification gaming arises when designers fail to consider the relevance of certain possibilities (e.g., that there are bugs in the game software) and therefore end up giving a formal problem specification which implicitly allows the undesirable solutions. But, as Krakovna et al. point out, the more complex the tasks become that ML systems are applied to, the harder it becomes for designers to consider every possible implication of their design

[3] See this link for a list of more than 60 examples collected by Krakovna et al. http://tinyurl.com/specification-gaming. Another example is the 'Proxy Problem', discussed by Johnson (forthcoming, fn 24). Here, an ML system uses a seemingly innocuous attribute (e.g., address) as a proxy for a protected attribute (e.g., race), which designers had deliberately excluded from the training data.

choices. Though it might be clear in retrospect how the unintended solutions could have been prevented, they can be highly surprising and difficult to predict in advance. Therefore, value transparency strategies focused on designers' motivations and choices would be unlikely to detect this type of value-ladenness.

These difficulties should of course not be used as an excuse for complacency. Rather, it puts even more onus on designers to carefully test and evaluate the impacts of their systems. Yet prospective value transparency still depends on our ability to recognise when something has (or preferably, will) go wrong. Again, the ability of advanced ML systems to find surprising solutions could potentially produce subtler impacts, or produce them through subtler causal chains, thereby making them more difficult to discover (explicitness) or explain (comprehensibility).

6. Conclusions

In this paper I have distinguished three strategies for managing value-ladenness in science—epistemic transparency, retrospective and prospective value transparency—and argued that these map onto different approaches within the literature on transparency in ML. I have argued that epistemic transparency faces limitations. However, so do both forms of value transparency.

Some constructive lessons can also be drawn. First, my discussion of transparency strategies in ML highlights the importance of distinguishing between retrospective and prospective value transparency strategies. More generally, it illustrates that the merits of transparency as a strategy for managing value-ladenness cannot be evaluated in the abstract.

The advantages and drawbacks of different strategies need to be investigated in specific contexts.

References

Angwin, Julia, Jeff Larson, Mattu Surya, and Lauren Kirchner. 2016. “Machine Bias.” *ProPublica*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Betz, Gregor. 2017. “Why the argument from inductive risk doesn’t justify incorporating non-epistemic values in scientific reasoning.” In *Current Controversies in Values in Science*, eds. Kevin Elliott, and Daniel Steel, 94–110. New York: Routledge.

Biddle, Justin. 2020. “On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning.” *Canadian Journal of Philosophy*, 1–21. doi:10.1017/can.2020.27

Brown, Matthew. 2020. *Science and Moral Imagination: A New Ideal for Values in Science*. Pittsburgh: University of Pittsburgh Press.

Burrell, Jenna. 2016. “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms.” *Big Data & Society*, June 2016. doi:10.1177/2053951715622512

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, Kevin. 2010. “Hydrogen Fuel-Cell Vehicles, Energy Policy, and the Ethics of Expertise.” *Journal of Applied Philosophy* 27:376–93.

Elliott, Kevin. 2017. *A Tapestry of Values: An Introduction to Values in Science*. Oxford: Oxford University Press.

Elliott, Kevin. 2020. “A Taxonomy of Transparency in Science.” *Canadian Journal of Philosophy*, 1–14. doi:10.1017/can.2020.21

Erasmus, Adrian, Tyler Brunet, and Eyal Fisher. 2021. “What is Interpretability?” *Philosophy & Technology*, 34:833-62.

Friedman, Batya, and Helen Nissenbaum. 1996. “Bias in Computer Systems.” *ACM Transactions on Information Systems* 14:330–47.

Gal, Yarin, and Zoubin Ghahramani. 2016. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.” *Proceedings of the 33rd International Conference on Machine Learning*, PMLR 48:1050-1059.

Johnson, Gabrielle. Forthcoming. “Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning.” *Journal of Moral Philosophy*.

Khosrowi, Donal. 2019. “Trade-Offs between Epistemic and Moral Values in Evidence-Based Policy.” *Economics and Philosophy* 35:49–78.

Kraemer, Felicitas, Kaes van Overveld, and Martin Peterson. 2011. “Is There an Ethics of Algorithms?” *Ethics and Information Technology* 13:251–60.

Krakovna, Victoria, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. 2020. “Specification gaming: the flip side of AI ingenuity.” *DeepMind Safety Research*, https://deepmindsafetyresearch.medium.com/specification-gaming-the-flip-side-of-ai-ingenuity-c85bdb0deeb4

Kroll, Joshua. 2018. “The fallacy of inscrutability.” *Philosophical Transactions of the Royal Society Part A* 376:20180084. http://dx.doi.org/10.1098/rsta.2018.0084

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.” In *Advances in Neural*

Information Processing Systems 30 (NIPS 2017), ed. I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. https://proceedings.neurips.cc/paper/2017

Mitchell, Tom. 1997. *Machine Learning*, New York: McGraw Hill.

Nguyen, Thi. 2021. Forthcoming. “Transparency is Surveillance.” *Philosophy and Phenomenological Research*, 1-31. doi:10.1111/phpr.12823

Passi, Samir, and Solon Barocas. 2019. “Problem Formulation and Fairness.” *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 39–48. https://doi.org/10.1145/3287560.3287567

Patel, Neel. 2017. “Why Doctors Aren’t Afraid of Better, More Efficient AI Diagnosing Cancer.” *The Daily Beast*, December 22 2017, https://www.thedailybeast.com/why-doctors-arent-afraid-of-better-more-efficient-ai-diagnosing-cancer

Obermeyer, Ziad, Brian Powers, Christine Vogelli, and Sendhil Mullainathan. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366:447–53.

Schroeder, Andrew. 2021. “Democratic Values: A Better Foundation for Public Trust in Science.” *British Journal for the Philosophy of Science* 72:545–62.

Selbst, Andrew, and Solon Barocas. 2018. “The Intuitive Appeal of Explainable Machines.” *Fordham Law Review* 87:1085–39.

Sullivan, Emily. 2022. “Understanding from Machine Learning Models.” *The British Journal for the Philosophy of Science* 73:109-33. doi:10.1093/bjps/axz035

Ward, Zina. 2021. “On Value-Laden Science.” *Studies in the History and Philosophy of Science* 85:54-62.

Zednik, Carlos. 2021. “Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence.” *Philosophy & Technology* 34:265–88.

Zednik, Carlos, and Hannes Boelsen. Forthcoming, “The Explanatory Role of Explainable Artificial Intelligence.” *Philosophy of Science*, http://philsci-archive.pitt.edu/18005/1/Zednik%20Boelsen%202020%20-%20Exploration%20and%20XAI.pdf

Zerilli, John. 2022. “Explaining Machine Learning Decisions.” *Philosophy of Science* 89(1):1-19. doi:10.1017/psa.2021.13