Inductive Risk, Understanding, and Opaque Machine Learning Models

forthcoming in *Philosophy of Science*

Emily Sullivan
Eindhoven University of Technology
Eindhoven Artificial Intelligence Systems Institute

Abstract:
Under what conditions does machine learning (ML) model opacity inhibit the possibility of explaining and understanding phenomena? In this paper, I argue that non-epistemic values give shape to the ML opacity problem even if we keep researcher interests fixed. Treating ML models as an instance of doing model-based science to explain and understand phenomena reveals that there is (*i*) an *external* opacity problem, where the presence of inductive risk imposes higher standards on externally validating models, and (*ii*) an *internal* opacity problem, where greater inductive risk demands a higher level of transparency regarding the inferences the model makes.

1. Introduction

Machine learning (ML) models can be so complex that it is unclear how the model arrived at its conclusions. The ML opacity problem inspired a proliferation of papers in computer science developing explainable AI (XAI) methods and philosophers describing various conceptions of opacity (Creel 2020). However, to what extent is opacity really a problem for explaining and understanding phenomena with ML models?

Recent work has provided some boundaries to the ML opacity problem. Creel (2020) suggests that depending on the type of opacity present—functional, structural or run opacity—there exists various XAI methods that make models transparent enough for the ends that researchers often have. Sullivan (2022) argues that going outside the model by reducing 'link uncertainty' (i.e. external evidential support connecting the model to the phenomena) is a more promising candidate for increasing understanding from ML models. Zednik (2021) claims the type of question a particular stakeholder asks—either a where, how, what or why question—matters for whether ML models are 'black boxes'.

In this paper, I provide more boundaries regarding the conditions under which ML opacity undermines explaining and understanding phenomena. I argue that the problem of model opacity is entangled with non-epistemic values. Particular ML use cases—medical diagnosis, tracking criminal activity, recommending which news to read—vary greatly in personal and social significance. It is not simply stakeholder interests or research questions that influence explanation, understanding, and ultimately the problem of opacity. I argue that even when we hold the target phenomena and researcher interests fixed, non-epistemic values impact when an ML model is capable of explaining target phenomena and the level of transparency necessary to enable understanding of phenomena.

My starting point is that ML modeling is another instance of doing model-based science. Thus, I am interested in the following traditional set of questions: How can models explain? When do models provide understanding of phenomena? When do models represent their targets? This starting assumption also implies that I am thinking of the ML opacity problem in a particular way: How much do we need to know about the model in order for it to explain, provide understanding, or say it accurately represents its targets? There are several other interesting questions about how ML models fit the paradigm of model-based science due to their predictive function, as well as ethical and political implications of model opacity (Burrell 2016). However, in this paper, I draw focus on the way that non-epistemic values impact the above set of traditional questions regarding model-based science. ML models are not necessarily unique in being entangled with non-epistemic values, but they do showcase the way that non-epistemic values have the potential to impact explanation, understanding, and general problems of model opacity.

I first introduce the problem of inductive risk and current applications to ML models (section 2). In section 3, I focus on an unexplored area of inductive risk impacting the ML modeling process: model acceptance. I then argue that inductive risk in ML model acceptance has consequences for the ML opacity problem for explanation and understanding (section 4). Non-epistemic values frame the problem of model opacity for explaining and understanding phenomena as: (*i*) an *external* problem, where higher standards on externally validating mozdsdels are necessary, and (*ii*) an *internal* problem, where the greater inductive risk demands a higher level of transparency regarding the inferences the model makes.

2. Inductive Risk and ML Models

There is always a risk of error with accepting (or rejecting) scientific hypotheses, theories, or using a scientific model in practice. A fundamental question in science is when to accept a given hypothesis in face of this risk and uncertainty. Proponents of inductive risk argue that hypotheses that have non-epistemic consequences that result from accepting (or rejecting) that hypothesis require the consideration of non-epistemic values (Rudner 1953; Steel 2013). Moreover, inductive risk seems to be present in several aspects of the scientific process beyond hypothesis acceptance, such as choosing a methodology, gathering and characterizing data, and the interpretation of data (Douglas 2000; Elliott 2013; Parker and Winsberg 2018). Tradeoffs between climate models, dosing for harmful chemicals, and tradeoffs between type I (accepting a false hypothesis) and type II errors (rejecting a true hypothesis) are common examples.

If the proponents of inductive risk are right, then we should expect that aspects of the ML modeling process are subject to inductive risk, with the necessary weighing of non-epistemic values at various stages in the ML pipeline. And indeed, there are yearly conferences like FAccT that focus exclusively on issues of fairness, bias, and values present in algorithms.[1] Mostly focus is placed on issues of bias, such as data biases that lead to different error rates for different populations or the way that certain architectures might exploit biases (Fazelpour and Danks 2021). Model transparency is seen as one possible solution to value-laden models and model biases. However, as proponents of inductive risk will argue, there is no value-free ideal, especially concerning algorithms (Johnson forthcoming). Yet there has been little philosophical engagement with how ML models are impacted by traditional considerations of inductive risk. Biddle (2020) is one exception, highlighting various aspects of the ML modeling process that are subject to ineliminable tradeoffs that reflect values, such as the tradeoffs exhibited in data choices and different conceptions of fairness that are impossible for a model to jointly satisfy and stressing that there is no value neutral way to construct ML models.[2]

While Biddle's discussion is insightful for thinking about value tradeoffs at several stages of constructing ML models, the way risk impacts issues of model justification or the grounds we have for accepting a ML model are conspicuously absent. It is not simply the *construction* of ML models that can arguably involve non-epistemic

[1] See: https://facctconference.org/

[2] See also Karaca (2021) for inductive risk in ML model construction.

values. After a model is constructed, we face the question whether to accept the model as being able to explain or provide understanding of particular phenomena and how non-epistemic values should influence accepting a model as having epistemic value. Getting closer to this goal, there has been some discussion on ML model choice. Dotan (2021) argues that the epistemic value of accuracy alone cannot be used to choose between computational models. However, Dotan considers aesthetic values, like simplicity, and does not explore social consequences or social values of model choice or acceptance.

In this paper, I explore how treating ML models as an instance of doing model-based science for explaining and understanding phenomena shows that ML model acceptance is subject to traditional questions of inductive risk concerning evidence and justification, and as a result, non-epistemic values encroach on the ML opacity problem.

3. Inductive Risk in ML Model Acceptance

In order for a model to be explanatory and provide understanding there must be a connection between the model and the target phenomena. On one common view, models explain and enable understanding when they capture patterns of counterfactual dependence that are true of their target (Bokulich 2011, Ylikoski and Kuorikoski 2010). Let's adopt this view as a working hypothesis for the following discussion. The first step is to identify a model's target or purpose. For example, a target could be answering a specific question, like whether there is racial bias in police stops in the United States. Once the target is identified, then assessing whether a model captures the counterfactual dependencies of interest in that target consists in two further checks for model adequacy: (*a*) whether the data the model is based on is adequate for establishing a true counterfactual dependence, and (*b*) whether the model, using adequate data, represents or finds counterfactual dependences true of the target. Both can be subject to inductive risk.

3.1. Accepting Data as Adequate for the Target Phenomena

As mentioned above, Biddle (2020) discusses how data choices that go into model construction are value-laden and reflective of non-epistemic values. Biddle often frames the discussion in terms of how (implicit) researcher interests and goals can influence data choices. However, researcher interest concerns identifying the *purpose* or target of the model (Parker 2020, Potochnik 2015).[3] Although the way non-epistemic values influence the direction of scientific inquiry is important, non-epistemic values still seem relevant once researchers identify a model target or purpose. Adopting the framework of inductive risk, I argue that non-epistemic values can be relevant for assessing whether, given a specific target or purpose, a particular dataset is adequate for establishing a true counterfactual dependence.

Consider the target of whether there is racial bias in police stops in the US. Pierson et al. (2020) took to examining racial disparities in police stops using a data approach. They discuss the importance of having the right data in order to evaluate whether there is racial bias in policing practices (i.e. assessing model adequacy based on (*a*)). One important data point—if not the most important for this purpose—is the race of the person stopped by police. In the data archives Pierson et al. accessed, race was recorded based on the stopping officer's perception. They argue that this method may introduce errors, suggesting alternative methods that potentially reduce these errors. In addition to self-reporting of the stopped individual, they suggest using a third-party's perception based on a driver's license photo (2020, 742).

In this example, we see that, in the opinion of the researchers, the model could more accurately find counterfactual dependencies in the target *if the dataset were constructed in a different manner*. The interests of the researchers do not change if they prescribe data methods that increase accuracy and potentially reduce error. The alternative data classification method is thought to capture the phenomena of racial bias more accurately, where one data method may (incorrectly) uncover racial bias that an alternative method would not. Inductive risk becomes relevant here. First, if the model suggests that police practices are *not* biased—and the model is wrong—this will prevent necessary public policy interventions. More related to (*a*), defining the epistemic concepts of 'error' and 'accuracy' and deciding which data establishes 'ground truth'—in the police bias case—involves the interplay with social and political values concerning racial identity. The practice of race labelling involves social / political considerations and has a history of marginalization and injustice (Browne 2015; Hanna et al. 2020; James and Burgos 2022). As Crawford and Paglen (2019) say, in struggles for justice, people have sought to "define the meaning of their own representations" and that "representations

[3] Bokulich and Parker (2021) discuss (*a*) in the context of their adequacy-for-purpose view applied to data models of highly physical systems, such as data collection from a Mars rover.

aren't simply confined to the spheres of language and culture, but have real implications in terms of rights, liberties and forms of self-determination."

Accepting that a particular dataset provides insight into racial bias requires adopting a particular definition of race which turns an epistemic question concerning model representation into one that is entangled with non-epistemic considerations that bear on how to define accuracy in a dataset and what kinds of error are acceptable. It is not simply a case of value tradeoffs or the balancing of false negatives and false positives; it is fundamentally about what data best captures the target. Thus, assessing a model based on (*a*)—the data is adequate for establishing a true counterfactual dependence—involves inductive risk and weighing non-epistemic values.

3.2 Accepting Dependencies as Representative of the Target Phenomena

Even if data classification issues are resolved there is still the question of (*b*): whether the model, using adequate data, represents or finds counterfactual dependences true of the target. Consider two more ML models: A melanoma detection and sexual orientation classification model. First, Esteva et al. (2017) developed a DNN that identifies cases of melanoma from healthy moles. The model was trained with semi-supervision of human labelled images of melanoma and healthy moles. When applied to a novel set of images, the model outperforms dermatologists at classification. Second, Wang and Kosinski (2018) developed a facial recognition model that seeks to identify the sexual orientation of individuals using DNNs. Briefly, this model uses roughly the same method as the melanoma model. The input data consisted of human labelled images of purportedly heterosexual men and women along with images of openly self-identifying gay men and lesbians. The model was said to be accurate at identifying sexual orientation when the model had five images of the same person. The researchers sought to add evidential support for the parental hormone theory (PHT), an origin theory for sexual orientation.

The targets in these cases are the visual appearance of a mole versus melanoma and how visual appearance could lend support to PHT theory, respectively. The high-level counterfactuals the models capture are: "If *x* and *y* visual patterns had not been present, then mole *M* would not have been a melanoma" and "if gender atypical facial topologies had been present, then sexual-orientation *O* would not have been heterosexual." Are the counterfactual dependencies found by the models true of their targets? Notice that the task of deciding whether a counterfactual dependence should be accepted as truly representative of a target becomes a traditional problem of inductive risk about how much evidence is necessary to accept or reject a hypothesis, and whether non-epistemic considerations impact the amount or kind of evidence required.

The visual appearance of a mole is the leading factor for medical intervention. The general types of counterfactuals the model provides already have significant medical support. There are, of course, risks involved with medical diagnosis that need to be considered when accepting whether the model should be used in practice. Traditional issues regarding the tradeoffs between type I and type II errors are relevant here. Moreover, because of bias in the training data, the model is only accurate on white skin and thus has generalization problems and should not be applied globally.

On the other hand, using facial appearance as getting at the truth of someone's sexual identity involves considerably higher demands for model acceptance beyond type I and type II errors. First, the dependencies the model finds between facial topologies and sexual orientation are scientifically controversial. Furthermore, as the proponents of inductive risk argue, social consequences that would result from accepting a hypothesis in error should be weighed. However, it can be difficult to weigh social consequences since the consequences themselves or the severity of them can be unknown. But in the case of phrenology and physiognomy we have considerable historical evidence of the kinds of consequences that have resulted from accepting such hypotheses by looking at its use in justifying racism, sexism, and eugenics. Given this track record, the amount of evidence needed to adopt a physiognomy-based theory is higher compared to accepting a competing theory. When considering Wang and Kosinski's model in particular, widespread discrimination throughout the world against gay individuals cannot be overlooked. Thus, the presence of such social implications creates a higher demand for evidence that the counterfactuals the model captures is true of the target. We better be reasonably sure that such a hypothesis is actually true before adopting it.

4. Inductive Risk and the ML Opacity Problem

How does the presence of inductive risk for accepting a model as being able to explain or provide understanding of phenomena impact the problem of ML opacity? The ML opacity problem emerges from model complexity, creating difficulty in understanding how the model arrived at its decisions. While it can be theoretically possible to document paths in a decision by tracing an input to an output, this documentation is often unhelpful.

Decision paths can be too complex for people to understand, and do not necessarily capture how all the subcomponents are related (Creel 2020). Solutions to opacity are relative to a given purpose or target, just like the adequacy of models is relative to a given purpose. As such, transparency is a stakeholder relative concept involving value tradeoffs about which aspects of the model need to be revealed in a human understandable way, and which aspects do not (Biddle 2020; Zednik 2021). Thus, non-epistemic values give shape the opacity problem through researcher capabilities and through identifying the explanatory targets of interest. However, even if we hold fixed the target and hold fixed the question a researcher wants answered, non-epistemic values influence the opacity problem in other ways. I argue that non-epistemic values influence both an external and internal problem for ML opacity.

4.1. Opacity as an External Problem

In previous work, I argued that the problem of model opacity is often not resolved by looking inside the model (Sullivan 2022). Instead, model opacity is an external problem connecting the model to the target; the problem of opacity is a function of how much '*link uncertainty*' the model has. The less evidence connecting the model to the target phenomena, the less understanding is possible from an opaque ML model. Among other examples, I used the above cases of the melanoma and sexual orientation classification models. The argument rests on the claim that these models seem to provide us with varying degrees of understanding while having similar amounts of model opacity. A better explanation for the varying understanding has to do with the level of independent empirical support connecting the model to the intended phenomena, instead of model opacity qua opacity. I did not consider how non-epistemic values might impact the problem of model opacity. However, if we accept there is inductive risk, in the link uncertainty framework it follows that non-epistemic values are entangled with the (external) problem of opacity.

Resolving the external problem of model opacity requires connecting the model's data and counterfactual inferences to the target phenomena, thereby reducing link uncertainty and defusing worries of opacity. Judgments about when there is enough evidence connecting a model to its target, such that model's links are strong *enough*, is target and domain specific. In cases where researchers are interested in whether counterfactual dependencies the model relies on are causally representative of phenomena, reducing link uncertainty will involve more traditional empirical research over and above the ML model.[4] In cases where the target of interest is establishing strong statistical correlations, reducing link uncertainty could alternatively require robustness checks with multiple models.

The greater the inductive risk, the higher demand there is for connecting the model's data and the model's inferences to the target, which amounts to a higher burden for reducing link uncertainty. Reducing link uncertainty can again range from improving ground truth methods for data collection to increasing statistical significance before accepting a counterfactual-inference as true of the target (Douglas 2000). For example, the higher threshold of evidence required to connect the sexual-orientation model to its target phenomena because of the risks involved, as discussed in section 3.2, directly entails a higher threshold of independent evidence is also necessary to overcome the model's opacity problem. If I am right that the extent to which model opacity poses a problem for explanation and understanding depends on the degree to which there is an external connection between the model and target, then the problem of opacity in ML is entangled with non-epistemic values since the process of accepting whether there is sufficient connection between the model and its target is itself entangled with non-epistemic values. Thus, exploring methods that externally validate models helps with overcoming inductive risk and opacity.

4.2. Opacity as an Internal Problem

Besides an external problem of opacity, inductive risk also exposes an *internal* problem for model opacity. The internal problem of model opacity requires verifying that extracted counterfactuals from the model are actually faithful to how the model works. In order to accept a model, there must be a set of counterfactual dependencies that we can extract from that model to measure against the target. In more traditional modeling methods—what Knüsel and Baumberger (2020) call process-based models—particular counterfactual dependences are deliberately built into the model. However, in data-driven ML models, counterfactual dependencies are not built-in, but must be extracted post-hoc. This is where ML model opacity strikes us as a problem. If the model is so complex that

[4] See Reichstein et al. (2019) for a hybrid method using ML models and physical models to understand climate phenomena.

it is unclear how the model makes its decisions, then how can we be sure that the counterfactual dependencies extracted are actually faithful to how the model works?

Various ML interpretability methods can provide some insight into the way models make decisions, such as feature importance methods that seek to identify the most salient features that determine a classification. Saliency maps, for example, highlight areas of an image that contribute most to the model's decision. In the melanoma detection model, a saliency map can highlight areas of the image with various pigmentation differences that the model relies on most for classification. Saliency maps can aid counterfactual reasoning about the target, along the lines of "If *x* and *y* visual patterns had not been present, then mole *M* would not have been a melanoma." The internal opacity problem of inductive risk requires a higher standard of verification that the counterfactuals extracted from the model are faithful to how the model makes its classifications, and that they are genuine correlations and not mere artifacts. A higher degree of model transparency is necessary. [5] Given the inductive risk present in the sexual-orientation model, more fine-grained counterfactual dependences about how the model made its decisions are required compared to the melanoma model or compared to a very benign model that, say, classifies handwritten numbers. And as it turns out, upon closer inspection of the sexual-orientation model, the researchers found—using saliency maps—the model places a high emphasis on superficial features, like makeup or hair style. Thus, the counterfactual inference the model made more often tracked "if *x* grooming pattern was *y*, then sexual-orientation *O* would not have been heterosexual." This suggests the model does not strongly rely on facial topologies, which was the primary target of interest. Thus, there is not only a higher burden of proof verifying that a dataset and a given counterfactual is representative of a target (the external problem), but there is a higher burden of proof for extracting the counterfactuals in the first place.

This is not to say that ML models with less inductive risk do not require verification. The claim is that models that have a higher level of inductive risk have a higher bar for verification. Moreover, we need to be aware of the limitations on current interpretability methods. For example, Wachter et al. (2018) provides an interpretability method that involves counterfactual extraction specifically, providing a counterfactual explanation along the lines of a "what if things had been different?" explanation. The

[5] I have in mind what Creel (2020) calls functional transparency, though non-epistemic values may also demand greater structural or run transparency.

method shows through hypothetical scenarios how small changes to certain feature values would result in changes to the outcome. However, this method has drawbacks, with the possibility to generate contradictory counterfactuals (Molnar 2019). Most, if not all, interpretability methods can be misleading to some degree. Saliency maps can be the same for multiple classes, calling their usefulness into question (Rudin 2019). Such limitations are the reason Rudin (2019) argues that opaque ML models should not be used in high stakes cases. Other models that have similar predictive accuracy—that are 'in principle' interpretable—should be used instead. [6]

To put the point differently, if a highly reliable ML interpretability method would require costly computational power compared to a less reliable method, models that have a high level of inductive risk would require researchers to opt for the costly method over the less costly method. So, while the current limitations on interpretability methods are not a barrier for explaining and understanding certain low stakes phenomena using ML models, the limitations can be a barrier for ML models that involve significant social implications. Thus, the presence of inductive risk not only has an impact on the external problem of ML opacity, but it also has an impact on the internal problem of model opacity. Accepting that a ML model is capable of explaining or providing understanding of phenomena requires a greater threshold for model transparency from models that face high stakes or face far reaching social consequences.

5. Conclusion

Most accept that non-epistemic values shape the problem of opacity insofar as non-epistemic values influence researcher or stakeholder interests. I argued here that even if we keep researcher interests fixed, there is a further sense in which non-epistemic values place a burden on model transparency. The higher the inductive risk, the greater demand there is to reduce link uncertainty and connect an ML model to its target through external validation (i.e. independent grounds connecting the model to the target) and the greater demand for internal model verification (i.e. extracting counterfactual inferences that are faithful to how the model works). Treating ML models as an instance of doing model-based science allows us to utilize helpful tools from the philosophy of science to ground discussions of ML model bias and their social and political implications.

[6] Perhaps model *use* in some high stakes contexts lessons the need for transparency (Durán and Formanek 2018). However, in this paper, I am concerned with explanation and understanding, not deployment.

Acknowledgements:

I'd like to thank the anonymous reviewers, my fellow symposium participates Justin Biddle, Rune Nyrup, and Kate Vredenburgh, Florian Boge, Thomas Grote, Kevin Kadowaki, Dunja Šešelja, Marianne van Panhuys, Philippe Verreault-Julien, and special thanks to John Mumm. This work is supported by the Netherlands Organization for Scientific Research (NWO grant number VI.Veni.201F.051) and the research programme Ethics of Socially Disruptive Technologies funded by the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

References:

Biddle, Justin B. 2020. "On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning." *Canadian Journal of Philosophy* 1-21.

Bokulich, Alisa. 2011. "How Scientific Models Can Explain." *Synthese* 180 (1): 33-45.

Bokulich, Alisa, and Wendy S. Parker. 2021. "Data Models, Representation and Adequacy-for-Purpose." *European Journal for Philosophy of Science* 11 (1): 1-26.

Browne, Simone. 2015. *Dark Matters*. New York: Duke University Press.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 1-12.

Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science* 87 (4): 568-89.

Crawford, Kate and Trevor Paglen. 2019. "Excavating AI: The Politics of Training Sets for Machine Learning." https://excavating.ai

Dotan, Ravit. 2021. "Theory Choice, Non-Epistemic Values, and Machine Learning." *Synthese* 198 (11): 11081-101.

Douglas, Heather. (2000). "Inductive Risk and Values in Science." *Philosophy of science* 67 (4): 559-79.

Durán, Juan M., and Nico Formanek. 2018. "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism." *Minds and Machines* 28 (4): 645-66.

Elliott, Kevin C. 2013. "Douglas on Values: From Indirect Roles to Multiple Goals." *Studies in History and Philosophy of Science Part A* 44 (3): 375-83.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542 (7639): 115-8.

Fazelpour, Sina, and David Danks. 2021. "Algorithmic Bias: Senses, Sources, Solutions." *Philosophy Compass* 16 (8): e12760.

Hanna, Alex, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. "Towards a Critical Race Methodology in Algorithmic Fairness." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 501-12.

James, Michael and Adam Burgos. 2022. "Race." *The Stanford Encyclopedia of Philosophy* ed. Edward N. Zalta, <https://plato.stanford.edu/archives/spr2022/entries/race/>.

Johnson, Gabbrielle. *Forthcoming*. "Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning." *Journal Moral Philosophy*.

Karaca, Koray. 2021. "Values and Inductive Risk in Machine Learning Modelling: The

Case of Binary Classification Models." *European Journal for Philosophy of Science* 11 (4): 1-27.

Knüsel, Benedikt, and Christoph Baumberger. 2020. "Understanding Climate Phenomena with Data-Driven Models." *Studies in History and Philosophy of Science Part A* 84:46-56.

Molnar, Christopher. 2019. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book.

Parker, Wendy S. 2020. "Model Evaluation: An Adequacy-for-Purpose View." *Philosophy of Science* 87(3): 457-77.

Parker, Wendy S., and Eric Winsberg. 2018. "Values and Evidence: How Models Make a Difference." *European Journal for Philosophy of Science* 8(1): 125-42.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff and Sharad Goel. 2020. "A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States." *Nature Human Behaviour* 4 (7): 736-45.

Potochnik, Angela. 2015. "The Diverse Aims of Science." *Studies in History and Philosophy of Science Part A* 53: 71-80.

Reichstein, Markus, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, and Nuno Carvalhais. 2019. "Deep Learning and Process Understanding for Data-Driven Earth System Science." *Nature* 566 (7743): 195-204.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1(5): 206-15.

Rudner, Richard. 1953. "The Scientist qua Scientist Makes Value Judgements." *Philosophy of Science* 20 (1): 1-6.

Steel, Daniel. 2013. "Acceptance, Values, and Inductive Risk." *Philosophy of Science* 80 (5): 818-28.

Sullivan, Emily. 2022. "Understanding from Machine Learning Models." *The British Journal for the Philosophy of Science* 73 (1): 109-33.

Wang, Yilun, and Michal Kosinski. 2018. "Deep Neural Networks are More Accurate than Humans at Detecting Sexual Orientation from Facial Images." *Journal of Personality and Social Psychology* 114 (2): 246-57.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology* 31 (2): 842–87.

Ylikoski, Petri, and Jaakko Kuorikoski. 2010. "Dissecting Explanatory Power." *Philosophical Studies* 148 (2): 201-19.

Zednik, Carlos. 2021. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy & Technology* 34 (2): 265-88.