How Values Shape the Machine Learning Opacity Problem

Sullivan, E. (2022). How Values Shape the Machine Learning Opacity Problem. In *Scientific Understanding and Representation* Eds. Lawler, Khalifa, Shech (pp. 306-322). Routledge.

Emily Sullivan
Eindhoven University of Technology
Eindhoven Artificial Intelligence Systems Institute

Abstract

One of the main worries with machine learning model opacity is that we cannot know enough about how the model works to fully understand the decisions they make. But how much is model opacity really a problem? This chapter argues that the problem of machine learning model opacity is entangled with non-epistemic values. The chapter considers three different stages of the machine learning modeling process that corresponds to understanding phenomena: (i) model acceptance and linking the model to the phenomenon, (ii) explanation, and (iii) attributions of understanding. At each of these stages, non-epistemic values can, in part, determine how much machine learning model opacity poses a problem.

1 Introduction

Machine learning (ML) models have an opacity problem. At least this is the impression that one gets by the 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? If we look to the built consensus that philosophers have taken on general issues of understanding and explanation, opacity is an insurmountable problem. Many agree that for a model to enable understanding, transparency (Dellsén 2020; Strevens 2013), simplicity (Bokulich 2008; Kuorikoski and Ylikoski 2015; Strevens 2008), and the ability to manipulate the model (De Regt 2017; Kelp 2015; Wilkenfeld 2013) are all necessary. But this cannot be the full story. While transparency is important, full-fledged transparency is not a sought-after goal. When we explain things, we often leave things out. Not all the details matter. Moreover, adding true but irrelevant or tangential details do not improve our understanding. Therefore, the important question is *what* should be transparent.

In this chapter, I argue that the problem of model opacity is entangled with non-epistemic values. I will look at three different stages of the scientific process surrounding ML models providing an understanding of phenomena: model acceptance and linking the model to the phenomenon (§2), explanation (§3), and attributions of understanding (§4). I argue that at each of these stages non-epistemic values, in part, determine how much ML model opacity poses a problem. My aim is to provide a broad outline about how non-epistemic values impact ML model opacity with regards to understanding and explanation. In the end, much more could be said about the role of non-epistemic values in each stage of the ML model pipeline. However, I hope that what I provide here illuminates that ML opacity, explanation, and understanding can be entangled with non-epistemic values.

2 Epistemic Risk in ML Model Construction and Acceptance

The role that non-epistemic values have in scientific theorizing and modeling practices has a rich history in the philosophy of science. Although there are those who argue that science should be a value free enterprise, there is a general consensus that values are ineliminable.[1] I will not rehearse these arguments here. Instead, I will argue that assuming the arguments that non-epistemic values are inescapable in scientific practice are successful—which I very much think they are—then non-epistemic values provide boundaries to the ML opacity problem.

Hempel (1965) and Rudner (1953) originally discussed *inductive* risk as the paradigm case of values entering scientific practice: there is always a risk of error when accepting whether a given hypothesis is true or false. When a hypothesis has downstream societal consequences, societal values should be part of its acceptance considerations (Douglas 2000). Others have since argued that there are other types of epistemic risk that require value choices beyond the risk of error in accepting (or rejecting) a scientific hypothesis (Biddle and Kukla 2017). For example, Harvard and Winsberg (2022) have recently argued that there are two core types of epistemic risk: inductive risk and representational risk. The former narrowly concerns the risk of error from endorsing a false hypothesis or statement, whereas the latter is the risk that a given scientific representation is inadequate for a given purpose. These kinds of epistemic risk are ever present among ML models.

The ML modeling requires considering tradeoffs between type I and type II errors, a standard tradeoff for inductive risk.[2] Moreover, Biddle (2020) pinpoints several aspects of the model pipeline that involves tradeoffs closer to representational risk that must be resolved in non-epistemic ways, such as identifying the problem to be modeled, training and benchmarking, algorithm design, and model deployment decisions. However, there is one specific area in the ML model pipeline that Biddle overlooks that is especially important when considering how much model opacity and complexity threatens explanation and understanding: model acceptance and establishing the link between the model and the phenomenon.

2.1 Epistemic Risk in Connecting ML Models to Phenomena

In a previous work, I argued that the problem of model opacity should not be understood as simply an internal problem that requires greater transparency of how the model works. Model opacity *qua* opacity need not undermine explanation or understanding from complex ML models. Instead, the problem of model opacity is largely an *external* problem connecting the model to the target. Specifically, the problem of opacity is a function of how much *link uncertainty* (LU) the model has (Sullivan 2022a). It is not the inner details of how the model works, but the higher-level abstract features that the model relies on to make its decisions, and most importantly, how those features are externally supported in providing insight into the target phenomenon that matters for understanding. There are various interpretability techniques available for "black-box" ML models that provides us with the necessary details regarding how the model made its decision such that the problem of model opacity becomes an external problem of LU.[3]

The framework of LU fits nicely with various theories regarding the way in which models provide understanding of phenomena. For example, a common view is that models explain when the counterfactual inferences that the model makes are true of their target (Bokulich 2011). Thus, one central aspect of accepting whether a model could be used to explain and enable understanding is in linking the model's inferences to the target phenomenon. In my view, when this link is weak or involves several uncertainties, understanding is limited. Moreover,

strengthening this link also dispels the problem of model opacity. On a different (yet arguably compatible) theory—the adequacy for purpose view (Parker 2020)—models are either adequate or inadequate for a very specific scientific purpose, such as answering a specific research question. Similarly, on such a view, the link between the model and the specific purpose needs to be established as adequate *enough* to provide insight into the research question identified. All said, one of the central features of accepting whether a model can provide understanding of phenomena is accepting whether the links between the model and the phenomenon are strong enough. Moreover, if the proponents of inductive risk are right, then deciding when we have reduced LU enough to understand can involve the consideration of non-epistemic values.

Consider an ML model that has a high degree of epistemic risk. Medical researchers sought to develop an accurate predictor model about the risk of death for patients presenting with pneumonia at a hospital. One goal of this model was to increase the efficiency of allocating medical resources to those who need it, while letting the others receive more comfort by recovering at home. Researchers found that an opaque neural network model achieved the highest accuracy rates. However, there is a clear epistemic risk in accepting whether the model should be used in practice, given the consequence of error.

It is my contention that there is also an epistemic risk in accepting that the model could *explain* or *provide understanding* of risk factors for patients with pneumonia. If scientists use the model to explain or represent the risks facing patients with pneumonia, but the model is inadequate for that purpose, then there are real non-epistemic consequences.[4] Moreover, the model explores the tradeoff between recovering from home and staying in the hospital and using up hospital resources, which is not purely epistemic. Thus, accepting whether a particular model is adequate to explain or provide understanding involves the (implicit) weighing of these values. This is a traditional problem of inductive risk (or representational risk) in deciding what kinds of evidence and the level of evidential support that is necessary in the face of uncertainty, and how strong a connection linking the model and the target is necessary.

2.2 Epistemic Risk and Opacity

If I am right, that the problem of model opacity is a function of LU, then the questions of inductive risk and representational risk are relevant for model opacity. Thus, ML models face *link uncertainty risk*. Judgments about when there is enough evidence connecting a model to its target, such that model opacity is not an epistemic barrier, can involve epistemic risk entangled with non-epistemic values.[5] Representational questions regarding what data should be used to represent the target phenomena, specific ML architectures suited for the problem, and even the specific interpretability technique chosen to gain high-level insight into black-box ML models require judgments that reflect values.

In the pneumonia case discussed above, as it turns out, when researchers sought to reduce LU, it became clear that the model did not provide understanding of the intended target of assessing which patients should be admitted. The data that the model used relied on the following underlying assumption:

> If a hospital-treated pneumonia patient has a very low probability of death, then that patient would also have a very low probability of death if treated at home. (Cooper et al. 1997, p. 136)

An especially astute observer may be able to see that such an assumption faces a large risk for inadequately representing the target. And indeed, a more interpretable rule-based model, trained on the same data, found that someone having asthma had a very low risk of death (see Caruana et al. 2015). However, the reason for the low risk of death was precisely because of the hospital treatment intervention. Patients with asthma are immediately placed in ICU care. The above representational assumption behind the opaque ML model faces not only a high degree of LU, but also a high degree of LU-risk because of the non-epistemic consequences. To my knowledge, current explainability techniques were not applied to the opaque ML model directly; instead, researchers inferred that the ML model likely made similar inferences as the interpretable model (Caruana et al. 2015). And thus, because of the potential downstream social consequences and the risk of misrepresentation, the ML model was not placed into practice, and researchers disregarded the epistemic value of the model.

Since the LU between the original neural network model and the target was high, the opacity of the model created a greater epistemic barrier. Moreover, since the LU-risk was high, the need for more research into the external connection between the model and the target increases further. Therefore, if the extent to which model opacity undermines explanation and understanding is based on the degree to which there is an external connection between the model and the 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.

3 Social Values and Explanation

Once researchers accept that a particular model is suitable for explaining phenomena, the next stage in the pipeline is actually constructing explanations. In this section, I argue that non-epistemic values, in part, determine the type and depth of the explanation that is required to adequately explain phenomena and the extent to which ML opacity poses an obstacle.

3.1 Non-Epistemic Explanatory Functions

Models on their own are not explanations; only when models help answer questions about some event or phenomenon do they explain (Bokulich 2011; Lawler and Sullivan 2021; Van Fraassen 1980). What I want to suggest here is that it is not just the specific question that we ask that matters for explanation, nor just the specific stakeholder or person who asks the question that matters (Zednik 2021). We must also consider the *function* or purpose of the explanation. Norms of explanation change depending on the function that an explanation has in a given context. Importantly, non-epistemic values are relevant when considering the functions that explanations should and do have, and the norms that follow.

The two explanatory functions that have gained the most attention in the epistemology and philosophy of science are the ontic and epistemic functions of explanation, namely, to discover relations in the world (Craver 2014; Illari 2013) and to enable understanding (Grimm 2010; Khalifa 2017). As a result, discussions about the norms of explanation are clustered around issues of representation (Frigg and Nguyen 2018), factivity (Elgin 2017), causality (Lange 2016; Sullivan 2019), and asymmetry (Reutlinger 2016). However, someone can explain for other purposes too. For example, an explanation is often sought to *justify* someone's actions. In this

case, a reasons explanation is warranted instead of the type of causal explanation often required for scientific explanation, which comes with its own norms (Majors 2007).

One notable difference between the way "explanation" is used in the computer science (CS) context compared with the philosophy of science is that, in CS, explanations are generally understood as a product separate from whatever model was used to make a decision or classification. This means that there are aims of explanation that are divorced from how the model itself works. For example, Tintarev and Masthoff (2007) discuss several different aims of explanations found in CS literature that are not epistemic, such as trust, effectiveness, persuasiveness, and satisfaction, among others. More recently, Lipton (2018) also discusses the different aims of recent XAI techniques, such as trust. Here too, depending on the aim or function of the explanation, the norms of what makes an explanation a *good* explanation change.

I will focus on two non-epistemic aims: trust and persuasiveness. Consider the example of an ad explanation on Amazon. On Amazon when you are searching for products to buy, the platform often provides the user with recommendations of additional products to look at and consider purchasing. These recommendations are generated by various types of ML models. Amazon provides the user with simple "explanations" explaining why they are seeing the recommendations that they do. The explanations are usually along the lines of: "because people who bought this product also bought this other one" or "sponsored products related to this item." Such explanations are built seamlessly into the platform so that users may not even realize that there is a question that needs answering. If the purpose of these explanations is for users to *feel* more trust toward the platform, then Amazon can conduct user studies to see whether the feeling of trust is actually increased. Do users trust Amazon more when this explanation is provided over this other one, or over not having an explanation at all? On the other hand, if the purpose is to persuade users to buy more products or to buy a certain product, then again, Amazon can measure the difference in user buying behavior just by changing the explanation. This is exactly the type of thing that platforms, like Amazon, do. The best explanation that satisfies these functions is an explanation that increases user trust or purchases. It does not matter whether the explanation is faithful to how the model works or satisfies other important epistemic norms to fulfill these functions. This means that model opacity is not a barrier to explaining if the purpose of the explanation is to build the impression of trust or to persuade.[6]

Discussing the function that ad explanations *should* have is beyond the scope of this chapter. However, if we assume that one of these functions is an epistemic function, such as to provide users with an understanding of how Amazon's recommendation algorithm works, then the above ad explanations provide little insight and fail to explain. In this latter case, we need more detail about how the model works, and the explanation would need to be true or at least *true enough* (Elgin 2017).

Discussions concerning the various functions that explanations can have, and the norms needed to satisfy these functions, demand social considerations. Even in the context of science, there are social factors that can determine the various epistemic functions of interest and how to satisfy these functions. Thus, non-epistemic considerations play a role in the explanatory phase of ML research insofar as researchers need to decide what purpose their explanations have, and some of these purposes are non-epistemic. Further, model opacity does not prevent explaining for various non-epistemic purposes.

3.2 How Non-epistemic Values Influence Epistemic Explanatory Purposes

What about cases where the purpose of an explanation is clearly an epistemic purpose, such as enabling understanding? Here too, non-epistemic values can impact the type of explanation that is required, what information is relevant, and the extent to which opacity is a problem.

First, when we are explaining various scientific phenomena, we often need to idealize some aspect of the phenomena to explain. Some phenomena are too complex to explain fully in an understandable way. Further, some argue that idealization improves an explanation even in the absence of complexity, because idealizations highlight the *difference-makers* in a way that a complete explanation does not (Strevens 2008). There is considerable discussion about what idealization norms entail (Weisberg 2007), and sometimes these norms depend on non-epistemic considerations. For example, Potochnik (2015, p. 76) argues that social aspects influence when something is *true enough* for explanation, even if we restrict the purpose of explanation to understanding. The research focus and context determine the way in which models can be idealized. For instance, researchers interested in explaining cooperative behavior could use the same evolutionary game theory model while focusing on different aspects, such as genetic differences, or non-selective traits, such as learning. It is largely the interests of scientists, and often the interests of funding bodies, that determine these research foci. I want to take Potochnik's discussion of social influences impacting explanation further beyond the research focus and interests of scientists. I want to suggest that the various interests of those who are *receiving* the explanation and the *practical domain* that the phenomenon is situated in impact the type of explanation, what information is relevant, and the extent to which model opacity is a problem.

Consider the COMPAS model (Northpointe 2012). It is a risk assessment model that uses ML technologies in determining risk for prison recidivism. It was developed by Northpointe (now Equivant), a profit-seeking private company. The model is opaque both in the sense that Equivant will not disclose the algorithm and because it is based on ML technologies. The COMPAS algorithm has been used in decisions regarding sentencing and parole in the United States. COMPAS has been charged with racial bias and using features such as a zip code to make its decisions (Angwin et al. 2016; Larson et al. 2016).

Consider, on the other hand, a different ML model that seeks to give a risk assessment about whether someone is at risk of developing certain types of cancers. Call this model HRisk+. Suppose further that this algorithm is also developed by a profit-seeking company that will not release its algorithm and it is opaque in the same way as COMPAS, because it is based on ML technologies. Further, suppose that HRisk+ is being used by doctors to decide whether certain patients should be considered for new medical trials or for increased medical screenings. HRisk+ also uses features such as a zip code to make its decisions.

Suppose further that it is the case that living in a particular zip code increases the risk that someone is arrested for a crime because of policing methods and living in the same zip code increases the risk of someone developing a particular type of cancer because it is an old Superfund site. Further, in both cases, there is an authority figure (judge or doctor) explaining why they came to a decision they did via an ML model. How do the non-epistemic differences in the COMPAS case and HRisk+ case change the requirements for explanation and proper model transparency? First, the *type* of explanation that is appropriate for why someone was denied parole or why someone was chosen for a medical trial differs because both the practical domain and the interests of the person receiving the explanation differ.

Consider the practical domain. In the case of the COMPAS model, the domain of interest is a sociopolitical domain. COMPAS is used solely to help determine whether an individual is

able to participate fully as a member of the larger social community. On the other hand, in the medical case, the domain of interest is the health sciences and clinical medicine. If a judge used the COMPAS algorithm to deny someone parole and gave an explanation simply citing the higher rates of crime and recidivism in the zip code in which that person lived, though perhaps true in the aggregate, this would not be satisfactory. The incarcerated person would rightly say, "it isn't relevant what someone *like me* in various respects might do, what matters is what *I* personally would do." However, in the medical case, the same type of statistical explanation that the judge provides seems completely appropriate, since medical decisions are a very different type of decision and are often based on aggregate patterns. In other words, in one case treating someone with, as King (2020) calls, the statistical stance is appropriate, but in another case it is not. To put this yet another way, it would not be surprising for a doctor to base a medical decision on what worked for your identical twin; however, it would be unjustified for a judge to make a sentencing determination based on what your identical twin did and not you. To be clear, it is not that the judge is wrong per se to use information regarding a zip code as evidence or even to use it in an explanation; however, given the sociopolitical domain in which the decision is situated, various non-epistemic considerations are equally or more salient, namely fairness and justice. Thus, what must be included in an explanation of why someone is being denied (or granted) parole is some connection between statistical trends and some *normatively* salient features that are relevant to the particular person under consideration and to larger norms of justice and fairness.[7] This is not the case with a medical explanation regarding who is a good fit for a medical trial or needs more screenings for a particular disease. Importantly, non-epistemic values are determining whether a statistical explanation over a reasons-based moral explanation is needed to explain the decisions of an ML model.

Further, the interests of those receiving an explanation can constrain the type of explanation that is required. For example, Zednik (2021) argues that various stakeholders in the ML modeling pipeline are interested in different epistemically relevant elements of how the model works. Zednik discusses this in terms of stakeholders asking different types of questions (e.g., where- vs why-questions) and that different question types require alternative levels of analysis. However, even if those receiving an explanation want an answer to the same broad question type—"why this decision?"—the specific interests of specific individuals can impact the explanation that is required. For example, someone who is a member of a group that is known to be subject to bias may be asking "why this decision?" specifically to find out whether bias was a part of the decision. The scope of such an explanation would include different epistemically relevant information compared with an explanation provided to someone who was not from such a group and where potential bias was not relevant. In the next section, I will discuss in more detail the role that different individual interests can have in impacting the scope of understanding.

The extent to which *model opacity* gets in the way of explanation also depends on the interests of the person receiving the explanation and its practical domain, and it thus depends on non-epistemic considerations. The level of detail needed concerning how the COMPAS model and the HRisk+ model work toward adequately explaining why a judge or doctor made a specific decision differs. Even though this is contested, it might be argued that in clinical settings, race, ethnicity, or gender are predictively useful (Vyas et al. 2020). For example, melanoma is a greater risk for white patients. However, in law, it is illegal in the United States and unjust for judges to make decisions regarding sentencing and parole based on race or to use various proxies for race in their decisions. Racial and gender bias in medical decisions looks different from criminal justice decisions. This suggests that the level of detail and transparency regarding how the model reaches a decision in the COMPAS case differs from the HRisk+ case, because the potential

problems of bias differ. Again, this difference is due to a difference in non-epistemic considerations.

All said, when we are explaining using ML models non-epistemic values partly determine the content of what makes for an acceptable explanation, and how much ML opacity poses a problem.

4 Opacity, Non-epistemic Values, and Attributions of Understanding

I briefly discussed two stages in the scientific process surrounding the use of ML models in providing understanding of phenomena—model acceptance and explanation—and how non-epistemic values, in part, determine the extent to which model opacity is a problem. Lastly, in this section, I consider the next stage: attributions of understanding from ML models. I argue that non-epistemic values also, in part, determine when ML model opacity prevents attributing understanding of phenomena.

4.1 The Stakes of Understanding

In epistemology, *pragmatic encroachment* theories of knowledge suggest that the stakes of a situation influence attributions of knowledge (Fantl and McGrath 2009; Hannon 2017). For example, if it is really important for someone to get to a meeting on time, that person may need more evidence of the train schedule than someone who does not have any particular place to be. I want to suggest that understanding from ML models also depends on the stakes. Specifically, depending on the stakes, someone might need to know more about how a given ML model works to understand.

It is not a settled question as to what constitutes understanding. Some argue that understanding is just a kind of knowledge (Riaz 2015), whereas others argue that understanding is distinct from knowledge (Hills 2016; Lawler 2019). I will not touch on this debate here. Instead, I aim at simply motivating that there are interesting cases in the ML context that suggest a pragmatic encroachment view of understanding. I aim at motivating these cases using common shared touchpoints for understanding.

One shared touchpoint for understanding is that understanding comes in degrees. Someone can understand something more or less. The simplest way to motivate a pragmatic encroachment view of understanding is to consider the more what-if-things-had-been-different questions someone can answer the *more* they understand (Hu 2019). In addition to describing the degree to which someone understands, we can, and should, talk about understanding attributions in terms of some minimum threshold condition.[8] On the simple view I am suggesting here, attributing understanding, or what we can call full-fledged understanding, depends on the threshold of the number of what-if questions in the set of all possible what-if questions on a given topic in a particular context that is necessary to attribute understanding. In some cases, if someone can only answer a few questions, then we should not attribute them with full-fledged understanding. For example, answering the simple question that the house burned down because of faulty wiring does not seem to be enough for *really* understanding why the house burned down. In some contexts, it seems that simply knowing the cause is too minimal for understanding. A fire marshal who is responsible for investigating the cause of the fire would surely need to answer more what-if questions for a proper attribution of understanding why the house burned down, such as why the wiring was faulty or the cause that sparked the wire failure. My claim here is

that if we accept that there is a minimum threshold for understanding attributions, then we can motivate a pragmatic encroachment view for when model opacity becomes a problem for understanding.[9]

There are two broad ways that non-epistemic values impact our attributions of understanding in the context of ML models: (*i*) the domain requires greater model transparency to attribute understanding, and (*ii*) the personal stakes in a given context can require greater model transparency to attribute understanding. I consider each in turn.

4.2 Varying Importance Concerning the Domain of Inquiry

First, consider how the domain of inquiry might demand greater model transparency for us to attribute an agent with understanding. A common way that recommendation systems work for various platforms is through a process called *collaborative filtering*. A collaborative filtering algorithm finds users that are similar to each other in various ways. It might be that they are in the same age group or that they tend to read the same news articles (e.g., sports and cryptocurrency). Various ML techniques are used to cluster similar users together.

Various domains use recommender systems based on collaborative filtering algorithms. However, different domains have more or less significance. For example, a news recommender system has lower stakes than a doctor recommender system. My claim is that there are different requirements for model transparency for understanding attributions based on the stakes of a given domain. In the news recommendation case, it is not important for the average user to know much at all about how users are clustered or how collaborative filtering works to attribute an understanding of why they are being shown a particular recommendation. An explanation along the lines of "users like you also enjoyed this article" seems sufficient.

However, in the case of a doctor recommender platform, given the importance that doctors have in someone's well-being (it could be a matter of life or death), a user would need to answer more what-if-things-had-been-different questions about how the algorithm works in order to understand. The consequences of error is greater in the doctor recommender case compared with the news recommender case. Because of the greater consequences of error, we need to be able to answer more what-if-things-had-been-different-questions regarding how the model works in order to attribute someone with understanding.

In some ways, a pragmatic encroachment view follows directly from the previous section that more model transparency is necessary in order to explain some phenomena compared with others. Given the close connection between explanation and understanding (Khalifa 2017; Strevens 2013), if explaining demands more details about how the model works, then it is necessary to know these more details to attribute understanding.

4.3 Greater Personal Stakes

Now, consider how the personal stakes of a practical situation could impact the demand for model transparency regarding the attributions of understanding. Imagine there are two people looking for a new apartment. Zoe needs an apartment quickly, whereas Thijs does not need one for at least 6 months or more. Zoe is also aware that recommendation platforms can have biases and make decisions based on race, gender, and nationality. Thijs, on the other hand, is not aware of such biases and does not fall into any of the concerned groups often impacted by bias. Both Zoe and Thijs are using a new housing platform that connects users to potential listings in their area.

The recommendation system is primarily designed using collaborative filtering technologies. Both Thijs and Zoe get recommendations and explanations that often include the phrase "users like you."

In order for Zoe to understand from these explanations, given the urgency of her particular situation, she needs to know more about what "users like you" means. Is the system filtering listings based on her nationality or her race? She needs to understand more about how the ML clustering algorithm works to gain understanding of why she is seeing the listings that she does. Specifically, she needs to know more about what makes a specific user *like her*. Is the recommendation system just using her preference profile for a home office and a children's playroom? Or is the recommendation system filtering out listings in specific neighborhoods because she comes from a country that is considered "non-Western"?

In other words, given how important it is for Zoe to find an apartment, and the higher risk of potential bias, she needs to be able to answer more what-if-things-had-been-different questions compared with Thijs for her to understand why she is seeing the recommendations that she sees. Greater model transparency is necessary to attribute understanding due to various non-epistemic factors regarding personal stakes.

5 Conclusion

One of the main worries with ML model opacity is that we cannot know enough about how the model works to fully understand the decisions they make. Without fully understanding how decisions are made how can we possibly trust the system or act based on its decisions? Everything I have said so far in this chapter is consistent with the view that there are some cases where the function of explanation is such, or the domain is such, or the stakes for an individual are such that the opaque nature of ML models prevents understanding. However, I have not argued for such a skeptical outcome across the board. Instead, my aim was to argue that non-epistemic factors contribute to the question as to how much of a problem opacity really is. I also have only focused on a cluster of issues surrounding explanation and understanding. There could be other reasons that demand greater model transparency, for example, being able to maintain privacy or some other value (Müller 2021). However, in order to explain and gain understanding with an ML model the problem of opacity greatly depends on features external to the model instead of features internal to it (i.e., link uncertainty and empirical support, explanatory functions, and the social and personal significance of the model and its domain).

Acknowledgments

For helpful comments and conversations, I would like to thank Thomas Grote, Insa Lawler, Elay Shech, and Mike Tamir, with special thanks to John Mumm. I presented this paper at the University of Rochester and I am very grateful for the conversation that resulted. This work is supported by the Netherlands Organization for Scientific Research (NWO grant number VI.Veni.201F.051). This work is also part of the research program Ethics of Socially Disruptive Technologies, which is funded by the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

References

Angwin, J., L. Jeff, M. Surya, and L. Kirchner. 2016. "Machine Bias." *ProPublica*. Accessed May 2021. https://www. propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Betz, Gregor. 2013. "In Defence of the Value Free Ideal." *European Journal for Philosophy of Science*, 3(2): 207–220. https://doi.org/10.1007/s13194-012-0062-x.

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

Biddle, Justin B., and Rebecca Kukla. 2017. "The Geography of Epistemic Risk." *Exploring Inductive Risk: Case Studies of Values in Science*, 215–237. https://doi.org/10.1093/acprof:oso/9780190467715.003.0011.

Bokulich, Alisa. 2008. *Reexamining the Quantum-classical Relation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511751813.

Bokulich, Alisa. 2011. "How Scientific Models Can Explain." *Synthese*, 180(1): 33–45. https://doi.org/10.1007/s11229-009-9565-1.

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." *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730. https://doi.org/10.1145/2783258.2788613.

Cooper, Gregory F., Constantin F. Aliferis, Richard Ambrosino, John Aronis, Bruce G. Buchanan, Richard Caruana, Michael J. Fine, et al. 1997. "An Evaluation of Machine-learning Methods for Predicting Pneumonia Mortality." *Artificial Intelligence in Medicine*, 9(2): 107–138. https://doi.org/10.1016/S0933-3657(96)00367-3.

Craver, Carl F. 2014. "The Ontic Account of Scientific Explanation." In A. Hutteman, M. Kaiser (eds.) *Explanation in the Special Sciences*, pp. 27–52. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-7563-3_2.

Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science*, 87(4): 568–589. https://doi.org/10.1086/709729.

Dellsén, Finnur. 2020. "Beyond Explanation: Understanding as Dependency Modelling." *The British Journal for the Philosophy of Science.* https://doi.org/10.1093/bjps/axy058.

De Regt, Henk W. 2017. *Understanding Scientific Understanding.* Oxford University Press. New York https://doi.org/10.1093/oso/9780190652913.001.0001.

Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science*, 67(4): 559–579. https://doi.org/10.1086/392855.

Elgin, Catherine Z. 2017. *True Enough*. MIT Press. Cambridge https://doi.org/10.7551/mitpress/9780262036535.001.0001.

Enoch, David, and Talia Fisher. 2015. "Sense and Sensitivity: Epistemic and Instrumental Approaches to Statistical Evidence." *Stanford Law Review*, 67: 557.

Fantl, Jeremy, and Matthew McGrath. 2009. *Knowledge in an Uncertain World*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199550623.001.0001.

Frigg, Roman, and James Nguyen. 2018. "The Turn of the Valve: Representing With Material Models." *European Journal for Philosophy of Science*, 8(2): 205–224. https://doi.org/10.1007/s13194-017-0182-4.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. "Explaining Explanations: An Overview of Interpretability of Machine

Learning." In 2018 *IEEE 5th International Conference on Data Science and Advanced Analytics* (DSAA), pp. 80–89. IEEE. https://doi.org/10.1109/DSAA.2018.00018.
Grimm, Stephen. 2010. "The Goal of Explanation." *Studies in History and Philosophy of Science Part A*, 41(4): 337–344. https://doi.org/10.1016/j.shpsa.2010.10.006.
Hannon, Michael. 2017. "A Solution to Knowledge's Threshold Problem." *Philosophical Studies*, 174(3): 607–629. https://doi.org/10.1007/s11098-016-0700-9.
Harvard, S., & Winsberg, E. 2022. "The Epistemic Risk in Representation." *Kennedy Institute of Ethics Journal 32*(1), 1-31. doi:10.1353/ken.2022.0001.
Hempel, Carl. 1965. *Aspects of Scientific Explanation.* Vol. 1. New York: Free Press.
Hills, Alison. 2016. "Understanding Why." *Noûs*, 50(4): 661–688. https://doi.org/10.1111/nous.12092.
Hu, Xingming. 2019. "Is Knowledge of Causes Sufficient for Understanding?" *Canadian Journal of Philosophy*, 49(3): 291–313. https://doi.org/10.1080/00455091.2018.1497923.
Illari, Phyllis. 2013. "Mechanistic Explanation: Integrating the Ontic and Epistemic." *Erkenntnis*, 78(2): 237–255. https://doi.org/10.1007/s10670-013-9511-y.
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. https://doi.org/10.1007/s13194-021-00405-1.
Kelp, Christoph. 2015. "Understanding Phenomena." *Synthese*, 192(12): 3799–3816. https://doi.org/10.1007/s11229-014-0616-x.
King, Owen C. 2020. "Presumptuous Aim Attribution, Conformity, and the Ethics of Artificial Social Cognition." *Ethics and Information Technology*, 22(1): 25–37. https://doi.org/10.1007/s10676-019-09512-3.
Khalifa, Kareem. 2017. *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press. https://doi.org/10.1017/9781108164276.
Kuorikoski, Jaakko, and Petri Ylikoski. 2015. "External Representations and Scientific Understanding." *Synthese*, 192(12): 3817–3837. https://doi.org/10.1007/s11229-014-0591-2.
Lange, Marc. 2016. *Because Without Cause: Non-Casual Explanations in Science and Mathematics*. Oxford University Press. New York https://doi.org/10.1093/acprof:oso/9780190269487.001.0001.
Larson, J., M. Surya, L. Kirchner, and J. Angwin. 2016. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*. Accessed May 2021. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
Lawler, Insa. 2019. "Understanding Why, Knowing Why, and Cognitive Achievements." *Synthese*, 196(11): 4583–4603. https://doi.org/10.1007/s11229-017-1672-9.
Lawler, Insa, and Emily Sullivan. 2021. "Model Explanation Versus Model-induced Explanation." *Foundations of Science*, 26(4): 1049–1074. https://doi.org/10.1007/s10699-020-09649-1.
Lipton, Zachary C. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue* 16, no. 3 (2018): 31-57.
Lusk, G. and Elliott, K.C., 2022. Non-epistemic values and scientific assessment: an adequacy-for-purpose view. *European Journal for Philosophy of Science*, *12*(2), pp.1-22.
Majors, Brad. 2007. "Moral Explanation." *Philosophy Compass*, 2(1): 1–15. https://doi.org/10.1111/j.1747-9991.2006.00049.x.

Müller, Vincent C. 2021. “Deep Opacity Undermines Data Protection and Explainable Artificial Intelligence.” *Overcoming Opacity in Machine Learning* 18.
Northpointe. 2012. COMPAS Risk and Need Assessment System: Selected Questions Posed by Inquiring Agencies. http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf.
Parker, Wendy. 2020. “Model Evaluation: An Adequacy-for-Purpose View.” *Philosophy of Science*, 87(3): 457–477. https://doi.org/10.1086/708691.
Potochnik, Angela. 2015. “The Diverse Aims of Science.” *Studies in History and Philosophy of Science Part A*, 53: 71–80. https://doi.org/10.1016/j.shpsa.2015.05.008.
Reutlinger, Alexander. 2016. “Is There a Monist Theory of Causal and Noncausal Explanations? The Counterfactual Theory of Scientific Explanation.” *Philosophy of Science*, 83(5): 733–745. https://doi.org/10.1086/687859.
Riaz, Amber. 2015. “Moral Understanding and Knowledge.” *Philosophical Studies*, 172(1): 113–128. https://doi.org/10.1007/s11098-014-0328-6.
Rudner, Richard. 1953. “The Scientist Qua Scientist Makes Value Judgments.” *Philosophy of Science*, 20(1): 1–6. https://doi.org/10.1086/287231.
Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Harvard University Press.
Strevens, Michael. 2013. “No Understanding Without Explanation.” *Studies in History and Philosophy of Science Part A*, 44(3): 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005.
Sullivan, Emily. 2019. “Universality Caused: The Case of Renormalization Group Explanation.” *European Journal for Philosophy of Science*, 9(3): 1–21. https://doi.org/10.1007/s13194-019-0260-x.
Sullivan, Emily. 2022a. “Understanding from Machine Learning Models.” *The British Journal for the Philosophy of Science*, 73(1): 109–133. https://doi.org/10.1093/bjps/axz035.
Sullivan, Emily. 2022b. “Inductive Risk, Understanding, and Opaque Machine Learning Models.” *Philosophy of Science, 1-13. doi:10.1017/psa.2022.62*
Tintarev, Nava, and Judith Masthoff. 2007. “A Survey of Explanations in Recommender Systems.” In 2007 *IEEE 23rd International Conference on Data Engineering Workshop*, pp. 801–810. IEEE. https://doi.org/10.1109/ICDEW.2007.4401070.
Van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford University Press. https://doi.org/10.1093/0198244274.001.0001.
Vyas, Darshali A., Leo G. Eisenstein, and David S. Jones. 2020. “Hidden in Plain Sight-Reconsidering the Use of Race Correction in Clinical Algorithms.” *New England Journal of Medicine*, 383(9): 874–882. https://doi.org/10.1056/NEJMms2004740.
Weisberg, Michael. 2007. “Three Kinds of Idealization.” *The Journal of Philosophy*, 104(12): 639–659. https://doi.org/10.5840/jphil20071041240.
Wilkenfeld, Daniel A. 2013. “Understanding as Representation Manipulability.” *Synthese*, 190(6): 997–1016. https://doi.org/10.1007/s11229-011-0055-x.
Zednik, Carlos. 2021. “Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence.” *Philosophy & Technology*, 34(2): 265–288. https://doi.org/10.1007/s13347-019-00382-7.

[1] See Betz (2013) for a defense of the value-free ideal and Johnson (forthcoming) for a defense of value-ladenness in science and algorithms.
[2] See Karaca (2021) for a discussion of type I and type II errors in ML and a proposal for cost-sensitive error classification to minimize inductive risk during model construction.

3 See Gilpin (2018) for a review of various ML interpretability methods.
4 See Lusk and Elliott (2022) for an adequacy-for-purpose theory of values in science.
5 In Sullivan [2022b], I call this the external problem of model opacity. I also discuss how values impact a further internal problem of opacity.
6 One worry here is that not every response to a why-question should count as an explanation. So, in what sense does an Amazon ad explanation aimed at persuasion actually count as an explanation at all, if it does not satisfy any epistemic constraints or have any real relationship to the ML model that it is meaning to explain? This objection already assumes that explanation has a specific function and thus comes with specific normative constraints. The suggestion here is that given that other scientific fields use the concept of explanation in a very different sense, instead of correcting their use of the concept, it is better to identify different functions and norms of explanation. Various functions of explanation can also be seen throughout philosophy, for example, causal, moral, mathematical, and metaphysical explanation.
7 For a discussion on statistical evidence in law, see Enoch and Fisher (2015).
8 See Kelp (2015), Khalifa (2017), and Wilkenfeld (2013) for discussions about degrees and thresholds for understanding.
9 Where exactly this threshold lies for understanding is beyond the scope of this paper; however, my argument relies on the claim that practical stakes have some influence.