Shaking up the dogma: Solving trade-offs without (moral) values in machine learning

Draft

Thomas Grote; Cluster of Excellence: "Machine Learning – New Perspectives For Science"; University of Tübingen; thomas.grote@uni-tuebingen.de.

Oliver Buchholz; Chair of Bioethics; ETH Zürich; oliver.buchholz@hest.ethz.ch

Abstract

The field of machine learning intricately links ethical and epistemological considerations in many contexts which raises the question as to their precise relation. This paper tries to provide a partial answer by focusing on one particular context, namely, the trade-off between accuracy and interpretability, which can be considered a prime example for the entanglement of ethics and epistemology in machine learning. At its core, the trade-off states that any choice of a machine learning model needs to balance the conflicting desiderata of achieving accurate predictions and an interpretable functionality. On a widely shared view inspired by the argument from inductive risk, this balancing of conflicting desiderata can only be resolved by appeal to non-epistemic values. By contrast, we argue that, in certain settings, the accuracy-interpretability trade-off can be resolved on purely epistemic grounds. To that end, we closely analyze the general nature of trade-offs as well as the notions of accuracy and interpretability. This allows us to derive strategies for resolving the accuracy-interpretability trade-off that center around choosing the right epistemic frame for a given machine learning application and, thus, do not require non-epistemic considerations. We conclude by sketching the implications of this result for the general relation of ethical and epistemological considerations in ML.

Keywords: machine learning; trade-offs; accuracy; interpretability; values in science:

1. Introduction

Ethical and epistemological issues are often enmeshed in machine learning. It is not uncommon for ethicists to wrestle with epistemic concepts like accuracy, reliability, opacity, or uncertainty alongside genuinely morally normative concepts like autonomy or fairness. Nevertheless, with few exceptions, the exact relationship between ethics and epistemology is rarely ever spelled out (Russo et al., 2024; Grote, 2024; Sterkenburg, 2024). This comes at the expense of an unclear understanding concerning the scope and methodology of ethics in machine learning. This paper is an attempt to make progress in this regard. Specifically, we study the relationship between ethics and epistemology on the basis of trade-offs in machine learning.

According to a widely held view, inspired by the literature on inductive and epistemic risk (see, e.g., Ward, 2021), machine learning models are value laden in that different trade-offs arise in the design and development process (Biddle, 2022; Nyrup, 2022; Johnson, 2023). These trade-offs amount to a choice between two desiderata that cannot be mutually satisfied. It is furthermore assumed that many of these trade-offs are inescapable: they occur necessarily and can only be resolved by means of value judgements, which then reflect in the model. If we accept these assumptions, then this results in a division of labor between ethicists and epistemologists in the design and development of machine learning models. On the one hand, the task of the epistemologist is to install appropriate safeguards so that the machine learning model meets a particular epistemic desideratum. On the other, the ethicist's task is to balance different value choices underlying certain trade-offs.

Even still, it is becoming apparent that the picture is more complicated than just suggested. Many trade-offs turn out to be false dogmas upon closer examination. For example, Beigang (2023) has argued that a trade-off between two statistical fairness notions, 'equalized odds' and 'predictive parity', can be modified by way of causal inference techniques so that they are universally compatible, whilst also retaining their intuitive appeal. Hence, the trade-off between the two fairness notions, of which the received view was that they are impossible for a machine learning model to satisfy simultaneously (Kleinberg et al., 2016), can be reconciled if machine learning developers use the right set of statistical techniques. This renders an ethical problem, which lies at the heart of the algorithmic fairness debate, into an epistemic.

Our main argument in this paper is that a similar story can be told about another trade-off, which is equally fundamental to the ethical debate, namely 'accuracy versus interpretability'. In its barest essentials, the trade-off states that when selecting a class of machine learning models, we are bound to choose between those that achieve high predictive accuracy but whose inner logic is opaque, or those whose inner logic is interpretable but whose predictive accuracy is inferior (when compared to the state of the art). Since sacrificing either accuracy or interpretability entails different epistemic risks, this is a consequential choice – particularly in high-stakes settings like healthcare (London, 2019). Contrary to this, we argue that if the deployment domain and the deployment purposes are well specified, the accuracy versus interpretability trade-off can be resolved: It is either possible to find a model that satisfies both desiderata or there is an epistemically clear preferable solution. The upshot is that the accuracy versus interpretability trade-off is first and foremost an epistemic as opposed to a morally normative issue as its resolution hinges on choosing the right epistemic frame for a given machine learning application.

While we are by no means the first to suggest that the accuracy- interpretability trade-off can be epistemically resolved (Rudin, 2019; Rudin et al. 2024), our approach represents progress in a twofold way: First, we make the current debate more nuanced by providing a detailed analysis of the structure of the trade-off that allows to identify multiple strategies for resolving it. Second, we generalize the current debate that primarily centers on the use of machine learning for social prediction (e.g., predicting the risk of recidivism or creditworthiness) by covering a broader range of applications, including vision-based tasks in healthcare or the use of machine learning for scientific discovery.

Accordingly, the paper proceeds as follows: Section 2 establishes conceptual common ground by developing an account of values in machine learning models. Section 3 tries to provide a precise definition of the accuracy versus interpretability trade-off by discussing the relevant philosophical and technical literature. We also look at historical precursors of the accuracy-interpretability trade-off like 'simplicity versus complexity' (Forster and Sober, 1994). In Section 4, we discuss different strategies to resolve the accuracy-interpretability trade-off. Finally, in Section 5 we draw general conclusions about the relationship between ethics and epistemology in machine learning based on our study of the accuracy versus interpretability trade-off.

2. Values in science and machine learning

Saying that science is not insulated from society and that scientific inquiry is encroached by non-epistemic values will not stir up much controversy these days. The starting point against the 'value-free ideal' in science has been the 'argument from inductive risk'. Initially, the argument states that in statistical testing, decisions about whether to accept or reject hypotheses should be informed by non-epistemic value judgments about the societal costs of accepting the hypothesis when it is false or rejecting it when it is true (Rudner, 1953; Douglas 2009; Steel, 2010). Over the last decades, however, the argument from inductive risk has undergone multiple conceptual expansions that

facilitate capturing a wider array of inductive/epistemic risks[1] that result from any choice regarding epistemic standards and methods that affect the acceptance or rejection of a hypothesis downstream (Biddle and Kukla, 2017).

In the same vein, the argument from inductive risk has been applied to various special sciences. Machine learning is the latest manifestation of this trend; and there are good reasons why it is particularly forceful in the case of machine learning: Just like statistical testing, machine learning is an inductive process, in which a model finds a function *f* to predict a variable of interest *Y* based on input data *X* within a given probability distribution *D*.[2] This entails that we cannot provide any *a priori* guarantees for the correctness of the output. Rather, the relevant guarantees are contingent on further conditions, such as *D* being sufficiently large and representative of the deployment setting (Johnson, 2023; Grote et al., 2024).

Biddle (2022) is arguably the most representative example of an epistemic risk approach to machine learning. Discussing the case of recidivism risk assessment, he challenges the supposed neutrality of machine learning models in arguing that, just like human decision-making, they are value laden: their design and development requires human decisions that, again, involve trade-offs reflecting human values. 'Values' can be best conceptualized as reasons that motivate or justify certain design choices in machine learning models (Ward, 2021). Moreover, although Biddle (2022) does not make this explicit, his usage of 'values' is tantamount to *non-epistemic* values, for instance, whose interests certain design choices serve and how to steer these design choices so that they lead to positive change (p. 322). We take up the distinction between epistemic and non-epistemic values later in this section.

He highlights these value choices by discussing how different trade-offs arise across the developmental cycle of machine learning models. Accordingly, model authorities need to navigate trade-offs, when (i) trying to operationalize the problem that the machine learning model is supposed to solve; (ii) selecting the training and evaluation data so that it is representative of the task at hand; (iii) balancing the model's accuracy versus interpretability; (iv) choosing fairness notions; (v) presenting outputs; and (vi) considering a wider range of transparency issues when the model is implemented into a given socio-technical environment. Importantly, even though Biddle is ultimately non-committal for which this is actually the case, the claim is that some of these trade-offs apply "in principle" and not just "in practice". That is, they are inescapable even in light of future research breakthroughs or resources (p. 324).

Many other philosophers follow the same playbook as Biddle. For example, Karaca (2021) uses the example of a binary classification model to detect cancer to highlight inescapable trade-offs in model construction and evaluation, such as choosing an appropriate performance metric that has to balance the epistemic risks of true positive and true negative instances. Here the model authorities must make social value judgments that take the epistemic risk profiles of the intended users into consideration. Nyrup (2022), in turn, discusses how different value decisions encroach the model design process and to what extent it is possible to make the relevant values transparent.

Against this backdrop, it is thus warranted to say that it has become the received view in the philosophy of machine learning that model authorities face various trade-offs throughout the developmental cycle, whereby they have to resort to *non-epistemic* value judgments in order to balance them. While we by no means are arguing for the strong claim that machine learning models are value-neutral (see also Phillips-Brown, 2023; Johnson, 2023; Sterkenburg, 2024), we understand

[1] Inspired by Biddle and Kukla (2017), we will just speak of 'epistemic risks' in the following, since the term is better able to account for a broader range of risks of error that potentially occur in the development cycle of machine learning models.

[2] Note, however, that according to Buchholz and Raidl (2022), the inductive component of machine learning is complemented by a pronounced falsificationist component.

our paper as a plea for more nuance: at least some of the purported trade-offs can be resolved solely on the grounds of *epistemic* values.

First of all, however, some conceptual clarifications are needed. Steel (2010) makes a useful proposal about how to demarcate epistemic from non-epistemic values. Central to his proposal is the assumption that epistemic values promote the acquisition of true beliefs. Note that epistemic values can be intrinsic or extrinsic. While intrinsic values are good in their own right, extrinsic values are a means for promoting intrinsic values. 'Accuracy' is the prime example of an intrinsic epistemic value, "because an empirically accurate theory is a theory whose consequences for observable phenomena are mostly true (p. 15)." In comparison, extrinsic epistemic values are characterized by the fact that they provide indirect support for the acquisition of true beliefs, as in the case of the testability of hypotheses, which enhances the efficiency of scientific inquiry. Non-epistemic values, by contrast, can be deemed to promote the attainment of moral or social goods.

Consider another term that is typically not defined in the literature, namely, 'trade-offs'. In essence, one can distinguish two conditions that are necessary for a trade-off to arise. According to the first,

> (T1) There must be a choice between two conflicting desiderata that cannot be satisfied simultaneously.

Take the example of a government organization that wants to evaluate policy interventions and decides to conduct a randomized controlled trial. Here the government organization has to balance a trade-off between the epistemic advantages of randomization (which is regarded as the best way to obtain unbiased estimates of the intervention's causal effects), and the resulting distributive justice problems (since one arm of the research participants are assigned to a seemingly inferior status quo policy) (MacKay, 2020).

In addition, according to the second necessary condition,

> (T2) Balancing the two desiderata must be a *hard choice* (Chang, 2017): Either alternative is better in some relevant aspects, and yet, neither seems to be as good as the other in all relevant aspects.

For example, if there were a study design available that offered the same epistemic advantages of randomized controlled trials without running into distributive justice problems, there would be no genuine trade-off. With that in mind, for present purposes, we are agnostic about the exact structural features of these hard choices – for instance, whether what makes trade-offs hard can be explained in virtue of ignorance of the normative and non-normative factors relevant to making the choice, whether the alternatives must be incommensurable, incomparable, or because the alternatives are on par (Chang, 2017).

The aim of this section was to discuss a widely shared view about the relationship between epistemic and non-epistemic values in machine learning. According to this view, non-epistemic values come into play when model authorities need to manage trade-offs occurring across the developmental cycle of machine learning models. Moreover, we introduced some key concepts from the values in science debate. On this basis, we now turn to a detailed analysis of the accuracy-interpretability trade-off.

3. The accuracy-interpretability trade-off

The trade-off between accuracy and interpretability claims center stage in the discourse on machine learning, since it relates two of the field's most basic concepts to one another. On an intuitive level, according to the trade-off, the choice of a class of machine learning models is inextricably linked to making a choice between models that achieve high predictive accuracy at the cost of an opaque functionality, or models whose functionality is interpretable at the cost of lower predictive accuracy.

In other words, "the most powerful machine learning techniques purchase […] predictive accuracy at the expense of our ability to access 'the knowledge within the machine'" (London, 2019, p. 15). Before discussing the different epistemic risks arising from this situation, let us take a closer look at the constituents of the trade-off and carve out what the concepts of accuracy and interpretability are (taken to be) about in this context.

Accuracy is arguably the less controversial concept involved in the trade-off, for it is a lot more intuitive to explicate and operationalize than interpretability. As a starting point, it is important to mention that in the context at hand, accuracy is commonly interpreted as *predictive* rather than *in-sample* accuracy, that is, as the accuracy that a machine learning model achieves on data that it did not have access to during the training process.[3] A standard way of evaluating this type of accuracy is by keeping a certain amount of data separate from the training data and let the trained machine learning model compute predictions for it that can subsequently be compared to the true values. Along these lines, (predictive) accuracy is usually defined as the loss that a machine learning model incurs on average over some set of data for which it issues predictions, where the loss measures the distance between the model's individual predictions and the corresponding true values (von Luxburg and Schölkopf, 2011).

Not only when compared to accuracy, interpretability is a considerably more elusive concept: it is hard to explicate, even harder to operationalize, and overall, it has been pointed out repeatedly that "it is not clear what it amounts to" (Räz, 2024, p. 159; London, 2019, p. 19). Nevertheless, attempts have been made to spell out what interpretability *might* amount to, acknowledging that the concept is domain-specific and very likely does not allow for an all-purpose definition (Räz, 2024; Rudin, 2019). For instance, Lipton (2018) distinguishes two broad meanings of the concept: a transparency notion of interpretability on the one hand and post-hoc interpretability on the other. According to the first meaning, interpretability concerns the extent to which the opaque functionality of machine learning models can be made transparent. Thus understood, interpretability provides insight into the mechanisms underlying machine learning models and sheds light on how they work. According to the second meaning, interpretability concerns the rationalization of model predictions 'after the fact', that is, after the model issued them. Thus understood, interpretability does not necessarily illuminate the precise functionality of machine learning models yet might nevertheless convey reasons for why a certain prediction was reached.[4]

While the post-hoc meaning of interpretability constitutes the rationale guiding research in the field of explainable AI, it is not the predominant view when it comes to defining the concept as such. Instead, there seems to be an emerging consensus that interpretability is related to the properties of machine learning models themselves.[5] Beginning with Rudin (2019) who argues that "an interpretable machine learning model is *constrained in model form* so that it is either useful to someone, or obeys structural knowledge of the domain" (p. 1), authors carved out a variety of properties that are deemed necessary for a model to be interpretable. For instance, Räz (2024) distinguishes two paradigms of interpretability with one concerning linear, and one concerning tree-based models. While this reasoning confirms once more that interpretability is not a monolithic concept, he carefully lists those properties that are needed in each of the paradigms such that a model is interpretable. In the case of linear models, this boils down to the particular form of the predictor function, which is simple and, thus, easy to grasp and work with for humans.

[3] The concept of predictive accuracy is thus closely related to what is discussed as the *test risk* in the technical literature, whereas in-sample accuracy corresponds to the *empirical risk* (von Luxburg and Schölkopf, 2011).

[4] A very different explication of the concept is due to Erasmus et al. (2021). On their account, interpretability is closely related to understandability and concerns a relation between explanations: through a process of interpretation, individuals can turn complicated explanations into ones that are more understandable.

[5] For instance, see Babic et al. (2021).

In the case of tree-based models, by contrast, interpretability does not hinge on the form of the predictor function, but rather on the partition of the input space in a geometrically simple way.

The last formulations already indicate how cashing out interpretability by appeal to specific properties of a machine learning model also connects the concept to considerations of simplicity. Briefly put, simpler models seem to be more interpretable. This is an interesting connection since just as interpretability, simplicity is known to be an epistemic desideratum that is in conflict with accuracy when selecting machine learning or, generally speaking, statistical models. Indeed, within the literature on statistical model selection, there is a decade-long debate concerning the trade-off between simplicity and accuracy (Glymour, 1980; Forster and Sober, 1994; Romeijn, 2017; Bonk, 2023). Independently of how simplicity is explicated, this trade-off centers around the impossibility of choosing a highly simplistic class of models and achieving high accuracy at the same time – either one has to sacrifice a certain degree of simplicity and fix a more complex class of models to achieve higher accuracy, or one has to sacrifice a certain degree of accuracy to end up with a simpler model. From a philosophical perspective, the debate about this trade-off is thorny since both simplicity and accuracy are typically considered to be methodological norms, that is, normative concepts that ought to be followed in the process of model selection and that are therefore in need of justification. However, while it is straightforward to justify accuracy as a methodological norm when the overarching goal consists in making accurate predictions, things become more difficult when it comes to justifying why one ought to choose simple models (Forster, 2002; Sterkenburg, 2025).

Exploring the details of this debate is beyond the scope of this article, yet the intimate connection between simplicity and interpretability and their shared incompatibility with accuracy highlight an interesting aspect about the two trade-offs: Whereas the justification of simplicity proves famously difficult, the justification of interpretability is in many cases even taken for granted without further mentioning, because it is deemed to be so obvious. Having access to more information about a machine learning model's inner workings, being able to manipulate it and analyze the consequences, or scrutinize the model's components seems *prima facie* beneficial after all (Lipton, 2018, p. 12). The intricate part about the trade-off between accuracy and interpretability, then, is not the justification of why the two concepts should be considered desirable in the first place. Instead, it is the fact that despite concerning two epistemic concepts, every way of settling the trade-off will lead to ethical ramifications. On the one hand, choosing a more interpretable class of models will lead to less accurate predictions (*pace* Rudin, 2019; Rudin et al., 2024). In a high-stakes setting like healthcare, this is a consequential choice, for it might lead to wrong or delayed diagnoses and, ultimately, treatment. On the other hand, choosing a less interpretable class of models to maximize accuracy might prevent human oversight and potentially violate existing legislation.

Consequently, it seems reasonable to assume that, in line with the argument from inductive risk, settling the trade-off requires value judgments, most importantly about how much accuracy or interpretability one is willing to sacrifice given the ethical implications of this choice. However, as we will point out in the next section, there are other ways of approaching the trade-off that sidestep such considerations and stay entirely in the epistemic realm.

Yet before moving on, note that while the accuracy versus interpretability trade-off is often taken for granted by philosophers and computer scientists (Huysmans et al., 2006; Dziugaite et al., 2020; London, 2019; Biddle, 2020), we are not aware of any actual proof that the very trade-off indeed exists. This is in contrast to related trade-offs like the one between certain statistical measures of algorithmic fairness. For instance, the trade-off concerning the incompatibility of equalized odds and predictive parity to determine the fairness of machine learning models has been formally proven on the grounds of an impossibility theorem (Kleinberg et al., 2016). Likewise,

Beigang's (2023) strategy to challenge the supposed incompatibility of said fairness notions borrows from Carnapian explication in that the respective fairness notions are re-engineered so that they are claimed to retain their intuitive appeal, whereby their compatibility is again proven mathematically.

By contrast, there are conceptual stumbling blocks to coming up with a proof for an accuracy versus interpretability trade-off. First, we are dealing with the combination of a narrowly defined (accuracy) and vague or at least context-dependent (interpretability) concept. Second, while accuracy can be considered to be an intrinsic epistemic value in that accurate predictions provide direct support for the acquisition of true beliefs, interpretability is an extrinsic, instrumental epistemic value: it enables model control, which, in turn, is conducive to the achievement of different epistemic ends, like justifying the model output, detecting biases, or reconciling the model output with human reasoning (Krishnan, 2020). We turn to this in detail in the next section and will, despite the lack of formal proof, stipulate for the time being that there exists a trade-off between accuracy and interpretability.

4. Resolving the accuracy-interpretability trade-off

The previous sections chartered the conceptual terrain: We laid down basic assumptions in the values in science literature, provided a definition of trade-offs in machine learning, and investigated the structure of the accuracy-interpretability trade-off. We can now use these conceptual tools to resolve the trade-off. Our strategy will be to stake out the logical space by analyzing a range of machine learning-based applications for the purposes of social prediction, medical diagnosis, or scientific discovery. It is important here that we also consider different model architectures. Based on this analysis, we will float the claim that if the domain and the task at hand are well-specified, there are two (mutually exclusive) ways out of the trade-off. Each of them targets one of the conditions establishing a trade-off introduced above: On the one hand, there are cases in which it is possible to make modeling choices so that the epistemic ends of accuracy and interpretability can be satisfied simultaneously, thereby circumventing condition (T1). On the other hand, there are cases in which one epistemic end clearly trumps the other, thereby circumventing condition (T2). In both scenarios, the context governing the trade-off between accuracy and interpretability are altered such that it is transformed into a methodological choice that either does no longer need to take into account the incompatibility of the epistemic ends in question or is no longer a *hard* choice in the sense of Chang (2017).

4.1 No way out: Eliminative strategies

Before outlining the two possible ways out of the accuracy-interpretability trade-off in greater detail, note that they differ considerably from what is commonly described as eliminative strategies in this context. These strategies explain away the components of the trade-off, such that there is nothing to resolve in the first place, simply because the trade-off does not exist – or so they argue.

Call the first such strategy *pragmatic eliminativism*. It is driven by empirical studies, most pertinent in human-computer interaction, that cast doubt on whether the purported epistemic benefits of interpretability translate into real-world settings (Poursabzi-Sangdeh et al., 2021; Bell et al., 2022; Kaur et al., 2024). In broad strokes, the methodology of these studies revolves around assigning research participants a cognitive task in which they are assisted by a statistical model. One group receives a simple model that is, say, linear and uses few variables, whereas the other group receives a complex black box model. It is then compared to what extent research participants are able to understand the model predictions, act upon the model predictions, or detect glaring errors in the model.

The results are brittle: For example, Poursabzi-Sangdeh et al. (2021) found that while research participants were able to better follow the model's predictions, this did not culminate in better decision-making when compared to the control group. Likewise, they were unable to correct for mistakes in the model. Kaur et al. (2024) found that research participants using interpretable models even became overconfident in the model's predictions, despite the output obviously being incorrect. If the findings of these studies are taken at face value, there is little point in trading off accuracy for interpretability – since the latter offers little epistemic value and, thus, ceases to be an epistemic end one should rationally strive for.

However, we plea for caution, as it is unclear in what way the design of the studies permit conclusions to be drawn for real world scenarios: The research participants are typically no experts and instead recruited via Mechanical Turk or from a student pool, the studies are either conducted online or in laboratory settings, and there are no proper incentives for research participants (Bell et al., 2022). More fundamentally, it is commonly hypothesized that the lack of instrumental epistemic value of interpretable models is owed to research participants facing information overflow. Yet, novelty effects are a possible confounder here, since the participants had no prior experience in using the model. As novelty effects wear off over time, it must be controlled in longitudinal studies if the problem of information overflow persists. That said, should we ever reach a point where several meta-analyses of externally valid studies show that interpretability has no instrumental epistemic value, then this would undermine any talk about an accuracy versus interpretability trade-off.

Consider a second strategy, which we call *conceptual eliminativism*. Its core idea is to critically analyze the concept of interpretability, leading to the conclusion that it is a vague and context-dependent notion. Based on this conclusion, it is often argued that one cannot meaningfully speak of an accuracy-interpretability trade-off before the concepts involved are clearly explicated or replaced with a set of desiderata and model properties that lend themselves more easily to operationalization (Krishnan, 2020; Lipton, 2018). In that sense, the trade-off is eliminated already on a conceptual level.

Leaving aside eliminative strategies that deny the existence of the accuracy-interpretability trade-off by explaining away its components, we will continue to stipulate that the trade-off indeed exists. However, as we briefly sketched above and will now argue in greater detail, there are at least two ways to proceed from this deliberately conservative starting point, even without any recourse to moral values.

4.2 A first way out: Achieving both desiderata simultaneously

The first way to proceed from acknowledging the existence of an accuracy-interpretability trade-off is based on the view championed by Rudin (2019) that one should always opt for what she calls inherently interpretable models. These are machine learning models relying on mathematical functions that have an intuitive, for instance, linear shape, incorporate domain-knowledge relevant to the particular application, include only a small number of meaningful features and are, thus, adequate means for achieving the epistemic end of interpretability. While this relation between certain model properties and interpretability may come as rather unsurprising, the second component of Rudin's approach is less obvious: Employing inherently interpretable models instead of complex black box models does not necessarily lead to a loss in accuracy as assumed by the accuracy-interpretability trade-off. Instead, "there is often no significant difference in performance between more complex classifiers […] and much simpler classifiers" (Rudin, 2019, p. 207). More recently, this point has been emphasized and formally investigated by Semenova et al. (2022) and Rudin et al. (2024). The bottom line of their analyses is that, indeed, there is a possibility of finding

simple-yet-accurate models in the sense that "for many problems, simple models can perform as well as much more complex models" (Rudin et al., 2024, p. 3).[6]

This immediately raises the question why one should care about the accuracy-interpretability trade-off if it is possible to achieve both epistemic ends at the same time: Doesn't the approach just outlined simply amount to another, third eliminative strategy that denies the existence of the trade-off altogether? Not quite. Upon closer inspection it becomes evident that the approach of achieving state-of-the-art predictive accuracy with inherently interpretable models is confined to specific settings. Indeed, note how in the quote above, Rudin et al. (2024) add the qualification 'for many problems' to their claim that interpretable and highly complex models can be on par with respect to their predictive performance. They even specify that these should be problems involving "tabular data" (p. 1) or, put differently, problems in which "the data are structured, with a good representation in terms of naturally meaningful features" (Rudin, 2019, p. 207).

As an illustration, Rudin (2019) uses a case study concerned with social prediction, namely, the widely discussed example of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). This is a complex proprietary machine learning model that is employed in the US justice system for predicting the risk of recidivism, that is, the probability that a defendant gets re-arrested after being released.[7] Pursuing the strategy of replacing complex black-box models with inherently interpretable ones, Rudin (2019, p. 209) proposes a simplistic model that is based on three decision rules and only involves 'age' as well as 'number of past crimes' as input features. Subsequently, she points out that this model "is equally accurate for recidivism prediction" (p. 209) as the original COMPAS model and, thus, achieves state-of-the-art accuracy as well as interpretability at the same time. Taking the methodological reasoning outlined above at face value, the reasons for this outcome should be obvious: The task of recidivism prediction gives rise to a setting in which, indeed, data are structured and have a good representation in terms of the two naturally meaningful features 'age' and 'number of past crimes'. Consequently, although accuracy and interpretability generally trade off against each other, the overall setup of this particular application, its *epistemic frame*, is such that both desiderata can be achieved at the same time.

Overall, then, the approach of using inherently interpretable models does not outright deny the existence of the accuracy-interpretability trade-off like the eliminative strategies outlined above. Instead, it generally acknowledges the existence of the trade-off and, for precisely this reason, emphasizes the necessity of running machine learning tasks in conditions that allow for accuracy and interpretability to be achieved simultaneously. This implies that the epistemic frame in which machine learning algorithms operate can be designed so as to circumvent condition (T1), thereby establishing one way of resolving the accuracy-interpretability trade-off.

At the same time, however, the exact conditions needed for this to work are a clear shortcoming of the approach. After all, it is well known that the benefits of using machine learning are largest in applications where the data is likely to be unstructured and features are not immediately meaningful. Thus, to resolve the accuracy-interpretability trade-off at scale, there has to be another way out of it that also works for the latter applications. Fortunately, there is one, as we shall argue next.

4.3 A second way out: Clearly preferring one desideratum to the other

[6] Note how the quotes in this paragraph illustrate the similarity between interpretability and simplicity discussed in Section 3.

[7] For details on COMPAS and its methodological and ethical ramifications, see Angwin et al. (2016), Larson et al. (2016) or Brennan et al. (2009).

So far, we have considered applications in which interpretable models can play to their strengths such that the accuracy-interpretability trade-off can be successfully mitigated by achieving both desiderata simultaneously. Yet there are more challenging edge cases.

One such case is the use of machine learning models for scientific discovery. AlphaFold and its iterations, showing remarkable success in the prediction of protein structures, represent the pinnacle here (Jumper et al., 2021; Abramson et al., 2024). And indeed, it seems obvious to point out that there are good reasons why, especially in later versions of the system, researchers relied on sophisticated transformer model architectures, as opposed to, say, an inherently interpretable decision-tree. Scientific discovery is also an area where the costs of the accuracy-interpretability trade-off become glaring, for it impedes the achievement of two fundamental goals, that is, prediction and understanding – the latter of which relates to the ability to explain the mechanisms that underlie phenomena such as protein folding.[8]

Even still, different strategies are available for resolving the trade-off in question. For instance, note that one natural interpretation of AlphaFold is that it is a throughout instrumentalist project: The goal is to predict increasingly complex protein structures as accurately as possible. These predictions can be deemed to be hypotheses about existing protein structures which then need to be further validated by means of experimental structure determination (Terwilliger et al., 2024).[9] Since we are not even close to understanding the mechanisms underlying protein folding, the goal of prediction trumps the goal of understanding for the time being. There is thus an epistemically clearly preferable solution in that one should select the model that promises the best predictive performance. This violates condition (T2) because, in this setting, deciding between accuracy and interpretability does no longer constitute a hard choice – which implies that there is no longer a genuine trade-off between both desiderata. Consequently, similar to the strategy above, the epistemic frame in which the machine learning model operates is set up such that there is a way out of the accuracy-interpretability trade-off, although here, condition (T2) rather than (T1) is targeted.

4.4 Diachronic strategies

In addition to the above, *diachronic* strategies, focusing either on condition (T1) or (T2), might as well help resolve the accuracy-interpretability trade-off in the context of scientific discovery. Note that, initially, it is typically complex and opaque models that pave the way for novel scientific breakthroughs, yet that we learn to understand over time which parts contribute to the outcome of interest in a meaningful way. This allows to develop simpler models that are more interpretable but roughly maintain the same level of predictive accuracy and, therefore, undermine condition (T1). Methodologically, various 'knowledge distillation' techniques (Hinton, 2015) form the basis here. We are aware that this approach has not yet been tested in practice on AlphaFold models, and that, in all likelihood, the translational process will prove to be intricate. Yet, at least in principle, this approach provides tools for resolving the accuracy-interpretability trade-off in scientific discovery.

Such a diachronic strategy for resolving the accuracy-interpretability trade-off naturally leads to another edge case, which is image-based diagnostics. Obviously, there are good reasons why both high predictive accuracy and interpretability are desirable in this case: On the one hand, incorrect diagnoses by the model can result in negative downstream effects for patients, which is why we should not sacrifice accuracy. On the other hand, it is crucial that physicians know when to trust and when to abstain from the model predictions (again, to avoid incorrect diagnoses). Interpretability therefore can contribute to aligning the reasoning of machine learning models and

[8] Our notion of scientific understanding is loosely based on Khalifa (2013).

[9] But see Zakharova (2024) for an opposing view.

physicians (Krishnan, 2020; Grote, 2023). Even still, just like for scientific discovery, it is typically unconstrained and ultimately opaque models that show the best predictive performance – at least under training conditions. Nevertheless, here too, we have to distinguish between the process of discovery and implementation. The models that perform best in benchmark competitions are not necessarily best suited for deployment in real-world clinical environments.

4.5 Hybrid strategies

Returning to the accuracy-interpretability trade-off, another strategy targeting (T1) is to adopt a hybrid approach in which interpretability is enabled by incorporating domain knowledge into a regular deep neural network. Chen et al. (2019) introduce the idea of a prototypical part network, where the model learns which prototypical parts are meaningful latent representations of an image. In addition, the model provides an activation map that visualizes the areas that the model uses to compute predictions and a confidence score that provides information on how (dis)similar the input image is to the learnt prototype. This approach has been applied by Barnett et al. (2021) for the classification of mass lesions in mammographic images. While the model's accuracy comes close to the state of the art[10], the benefits with respect to interpretability are two-fold: it constrains the reasoning style of the model in a way that aligns with those of physicians, while also making the reasoning process intelligible to them.[11]

Importantly, while hybrid strategies[12] allow for resolving the accuracy-interpretability trade-off on purely epistemic grounds, it has to be pointed out that they rely on a subtle semantic shift concerning the notion of interpretability. As the models of choice are deep neural networks, we give up on the idea that interpretability is intimately linked to simplicity or linearity (Räz, 2024), but the built-in domain knowledge becomes the distinctive factor instead.

In sum, then, one can distinguish between two broad strategies that acknowledge the existence of the accuracy-interpretability trade-off and propose a way out of it. Both are based on the idea of purposefully designing the epistemic frame in which a machine learning model operates to undermine the conditions establishing the trade-off. The first strategy is targeting condition (T1). It consists in providing structured data and meaningful features or specific domain-knowledge that allow interpretable models to achieve state-of-the-art predictive performance. The second strategy is targeting condition (T2). It consists in deploying machine learning models in settings where accuracy clearly trumps interpretability or *vice versa*. In the following section, we will discuss implications of this state of affairs with a view on the relationship between ethics and epistemology in machine learning.

5. Revisiting the relationship between ethics and epistemology in machine learning

[10] Note, however, that any statements considering state-of-the-art performance in machine learning are inevitably contingent on the time of testing and the respective benchmark task.

[11] Improving the alignment between machine and human judgment also opens the possibility of a socio-technical approach to overcoming the accuracy-interpretability trade-off. The basic idea here is that even if an interpretable machine learning model were to perform below state-of-the-art, interpretability results in further epistemic advantages, so that as a tandem, the accuracy of a human expert plus the interpretable machine learning model surpasses the accuracy of a more powerful but opaque model. See Fazelpour (2024) for a general discussion about how a socio-technical approach can be used to resolve various trade-offs that play a prominent role in the ethics of machine learning discourse. In contrast to this, however, our paper focusses on model-centric strategies for how to cope with trade-offs. Moreover, for reasons discussed in Section 4.1, we have some cautious skepticism about whether the feasibility of this approach is sufficiently supported by empirical studies.

[12] See also Ilanchezian et al. (2021) for a similar-minded approach, revolving around BagNet models.

Progress in philosophy is arguably achieved through an increased understanding of the network of dependence relations between phenomena (Dellsén et al., 2024). Our goal in this paper has been to elucidate the dependence relations between moral and epistemic values in machine learning. Looking at the debate through a historical lens – and although, admittedly, the debate is barely more than five years old – it was initially the importance of non-epistemic values that has been stressed in the relevant literature (Biddle, 2022; Johnson, 2023). Lagging behind, by comparison, has been the understanding of what issues exactly fall within the scope of epistemology. But slowly, the tides are starting to turn.

For example, Russo et al. (2023) paint a nuanced picture concerning the way of how moral values are entangled with epistemic values in machine learning. Ultimately, they argue for a holistic process of model development and validation, in which a mere quantitative assessment is supplemented by a reflective praxis concerning the moral values that must be embedded in a given model to ensure its trustworthiness. Moreover, Ratti and Russo (2024) emphasize the bidirectionality of epistemic and moral values concerning the development and deployment of machine learning models: While reflections on epistemic risks can guide the design and assessment of machine learning models, their deployment can also disrupt the moral norms in a domain of interest, such as healthcare or criminal justice. Both papers take a macro-perspective on the entanglement between moral and epistemic values in machine learning and can be deemed to be conceptual expansions of existing work from the 'values in science' literature and Science and Technology Studies.

While we agree with these accounts for the most part, we see our paper as part of an emerging literature that, at a micro-level, argues for *methodological prioritanism* of epistemology over ethics. As we pointed out at length above, it turns out that some issues like the accuracy-interpretability trade-off where epistemic and moral values seem to be intertwined can be best tackled by making appropriate epistemic modeling choices. Aside from the aforementioned paper by Beigang (2023) that seeks to resolve the trade-off between different fairness notions through a combination of conceptual re-engineering and causal inference techniques, Sterkenburg (2024) argues that it does not follow from the argument from inductive risk that learning algorithms (but not the trained model), referring to the inferential process mapping the input data to the output, must be necessarily laden with moral values. This lays the ground for a more fine-grained charting of the interactions between moral and epistemic values. Grote (2024), again, develops a revisionary account that reframes many of the key issues in the ethical debate about machine learning in healthcare as epistemic problems. To illustrate, rather than worrying about how the involvement of machine learning models may diffuse the attribution of responsibility in the case of diagnostic errors, the claim is that the best mitigation strategy involves installing appropriate epistemic guardrails so that the epistemic authority of clinicians is being preserved. Among others, this requires establishing epistemic norms for how to aggregate machine and human judgment, or how to resolve disagreements in light of information asymmetries. Finally, Zezulka and Genin (2024) argue that when algorithmic fairness questions are reconceptualized as policy problems that seek to anticipate the impact of the model output on the allocation of social goods, formal fairness notions all have undesirable consequences. Drawing on the causal inference toolkit, they highlight that a clearly preferable alternative instead is to model the counterfactual treatment outcomes of a given policy.

Although the account provided in this paper falls squarely into the abovementioned line of research, we think that our contribution goes beyond adding just another piece of mosaic to the overall picture. Most notably, we provide a general methodology for dealing with trade-offs in machine learning, which, in its barebones, consists of the following steps: (i) Explicate the desiderata involved in the trade-off and specify its underlying structure; (ii) specify the epistemic

and moral values related to the conflicting desiderata; (iii) map out a wide range of deployment scenarios; and (iv) consider what modelling choices allow for resolving/relaxing the trade-off for different deployment scenarios. Even if it turns out that in some cases, there may be no epistemic solution available, this approach leads to a better understanding of when exactly we have to resort to moral values to cope with trade-offs.

With that in mind, our account also has limitations. One is that our analysis of the accuracy-interpretability trade-off has been confined to instances of supervised learning. Hence, we did not consider how this trade-off arises and can be resolved for the latest generation of machine learning models, called 'foundation models'. These are trained on broad data at a massive scale via self-supervised learning, resulting in models that are able to perform various downstream tasks – most pertinently natural language processing. The main reason for not addressing the accuracy-interpretability trade-off concerning foundation models is that even accuracy, the desideratum which we have basically taken for granted in this paper is in shambles: Since foundation models are commonly trained on *the entire internet*, whilst lacking a clear data source, it is unclear how to discern generalization from memorization capacities, or how to determine the boundary conditions for a model's predictive accuracy (Grote et al., 2024). More research on the theoretical underpinnings of foundation models is thus necessary before the accuracy-interpretability trade-off arising in this context can be tackled in a meaningful way.

6. Conclusion

This paper made an attempt to clarify the relation between ethics and epistemology in machine learning by closely analyzing one of the field's central trade-offs, the one between accuracy and interpretability. In contrast to the widely shared view inspired by the argument from inductive risk that trade-offs in machine learning can only be resolved by appeal to non-epistemic values, our analysis reveals that this need not always be the case. More precisely, we carved out two general strategies for resolving the accuracy-interpretability trade-off through carefully designing the epistemic frame of a given machine learning application and, thus, by purely epistemic means. The first strategy consists in providing structured data and meaningful features or specific domain-knowledge that allow interpretable models to achieve state-of-the-art accuracy, thus making both desiderata achievable simultaneously. The second strategy, in turn, consists in deploying machine learning models in settings where accuracy clearly trumps interpretability or *vice versa*, thus making one desideratum clearly preferable to the other.

In sum, the present paper suggests that, at least under specific conditions, epistemic considerations are all that is needed to settle trade-offs in machine learning. This result ties in well with recent literature that points into a similar direction by (explicitly or implicitly) arguing for a methodological priority of epistemology over ethics in machine learning (Beigang, 2023; Grote, 2024). Nevertheless, while contributing to this strand of research, the present paper also highlights several gaps, such as the questions of whether and to what extent our results could be generalized to foundation models, other trade-offs, or even other issues in machine learning altogether that also include ethical and epistemological considerations. Answering these will be the subject of future work.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1-3.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica; 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science*, 373(6552), 284-286.

Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., & Rudin, C. (2021). A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12), 1061-1070.

Beigang, F. (2023). Reconciling Algorithmic Fairness Criteria. *Philosophy & Public Affairs*, 51(2).

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 248-266).

Biddle, J. B. (2022). On predicting recidivism: epistemic risk, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy*, 52(3), 321-341.

Biddle, J. B., & Kukla, R. (2017). The geography of epistemic risk. In Elliott, K. & Richards, T. (eds.): *Exploring inductive risk: Case studies of values in science*, pp. 215-237. Oxford University Press.

Bonk, T. (2023). Functionspaces, Simplicity and Curve Fitting. *Synthese*, 201(2):58.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36(1), 21-40.

Buchholz, O., & Raidl, E. (2022). A falsificationist account of artificial neural networks. *The British Journal for the Philosophy of Science*, online first, DOI: https://doi.org/10.1086/721797

Chang, R. (2017). Hard choices. *Journal of the American Philosophical Association*, 3(1), 1-21.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Dellsén, F., Firing, T., Lawler, I., & Norton, J. (2024). What is philosophical progress? *Philosophy and Phenomenological Research*, 109(2), 663-693.

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Dziugaite, G. K., Ben-David, S., & Roy, D. M. (2020). Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. arXiv preprint arXiv:2010.13764.

Erasmus, A., Brunet, T. D., & Fisher, E. (2021). What is interpretability? *Philosophy & Technology*, 34(4), 833-862.

Fazelpour, S. (2024). Disciplining deliberation: a sociotechnical perspective on machine learning trade-offs. arXiv preprint arXiv:2403.04226.

Forster, M. R. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of Science*, 69(S3), S124-S134.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1), 1-35.

Glymour, C. (1980). *Theory and Evidence*. Princeton University Press.

Grote, T. (2023). The Allure of Simplicity: On Interpretable Machine Learning Models in Healthcare. *Philosophy of Medicine*, 4(1).

Grote, T. (2024). Machine learning in healthcare and the methodological priority of epistemology over ethics. *Inquiry*, 1-30.

Grote, T., Genin, K., & Sullivan, E. (2024). Reliability in machine learning. *Philosophy Compass*, 19(5), e12974.

Grote, T., Freiesleben, T., & Berens, P. (2024b). Foundation models in healthcare require rethinking reliability. *Nature Machine Intelligence*, 6, 1421-1423.

Hinton, G. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

Huysmans, Johan and Baesens, Bart and Vanthienen, Jan, Using Rule Extraction to Improve the Comprehensibility of Predictive Models (2006). K.U. Leuven KBI Working Paper No. 0612, URL: http://dx.doi.org/10.2139/ssrn.961358

Ilanchezian, I., Kobak, D., Faber, H., Ziemssen, F., Berens, P., & Ayhan, M. S. (2021). Interpretable gender classification from retinal fundus images using BagNets. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2021* (pp. 477-487). Springer International Publishing.

Johnson, G. M. (2023). Are algorithms value-free?: Feminist theoretical virtues in machine learning. *Journal of Moral Philosophy*, 1(aop), 1-35.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.

Karaca, K. (2021). Values and inductive risk in machine learning modelling: the case of binary classification models. *European Journal for Philosophy of Science*, 11(4), 102.

Kaur, H., Conrad, M. R., Rule, D., Lampe, C., & Gilbert, E. (2024). Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 1-34.

Khalifa, K. (2013). The role of explanation in understanding. *The British Journal for the Philosophy of Science*, 64(1), 161-187.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

Krishnan, M. (2020). Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3), 487-502.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica; 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31-57.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, 49(1), 15-21.

MacKay, D. (2020). Government policy experiments and the ethics of randomization. *Philosophy & Public Affairs*, 48(4), 319-352.

Nyrup, R. (2022). The limits of value transparency in machine learning. *Philosophy of Science*, 89(5), 1054-1064.

Phillips-Brown, M. (2023). Algorithmic neutrality. arXiv preprint arXiv:2303.05103.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-52).

Ratti, E., & Russo, F. (2024). Science and values: a two-way direction. *European Journal for Philosophy of Science*, 14(1), 6.

Räz, T. (2024). ML interpretability: Simple isn't easy. *Studies in History and Philosophy of Science*, 103, 159-167.

Romeijn, J.-W. (2017). Inherent Complexity: A Problem for Statistical Model Evaluation. *Philosophy of Science*, 84(5):797–809.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.

Rudin, C., Zhong, C., Semenova, L., Seltzer, M., Parr, R., Liu, J., ... & Boner, Z. (2024). Amazing things come from having many good models. arXiv preprint arXiv:2407.04846.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 20(1), 1-6.

Russo, F., Schliesser, E., & Wagemans, J. (2024). Connecting ethics and epistemology of AI. *AI & Society*, 39(4), 1585-1603.

Semenova, L., Rudin, C., & Parr, R. (2022). On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1827-1858).

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science*, 77(1), 14-34.

Sterkenburg, T. F. (2024). Values in machine learning: What follows from underdetermination?. PhilSci Archive preprint, URL: https://philsci-archive.pitt.edu/24439/

Sterkenburg, T. F. (2025). Statistical Learning Theory and Occam's Razor: The Core Argument. *Minds and Machines*, 35(1), 1-28.

Terwilliger, T. C., Liebschner, D., Croll, T. I., Williams, C. J., McCoy, A. J., Poon, B. K., ... & Adams, P. D. (2024). AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods*, 21(1), 110-116.

von Luxburg, U., & Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In Gabbay, D.M., Hartmann, S., & Woods, J. (eds.): *Handbook of the History of Logic* (Vol. 10), pp. 651-706. North-Holland.

Ward, Z. B. (2021). On value-laden science. *Studies in History and Philosophy of Science Part A*, 85, 54-62.

Zakharova, D. (2024). The Epistemology of AI-driven Science: The Case of AlphaFold. PhilArchive preprint, URL: https://philsci-archive.pitt.edu/24182/.

Zezulka, S., & Genin, K. (2024). From the Fair Distribution of Predictions to the Fair Distribution of Social Goods: Evaluating the Impact of Fair Machine Learning on Long-Term Unemployment. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1984-2006).