Are Algorithms Value-Free?

Feminist Theoretical Virtues in Machine Learning

Gabbrielle M Johnson

Forthcoming in *Journal of Moral Philosophy*

Abstract

As inductive decision-making procedures, the inferences made by machine learning programs are subject to underdetermination by evidence and bear inductive risk. One strategy for overcoming these challenges is guided by a presumption in philosophy of science that inductive inferences can and should be value-free. Applied to machine learning programs, the strategy assumes that the influence of values is restricted to data and decision outcomes, thereby omitting internal value-laden design choice points. In this paper, I apply arguments from feminist philosophy of science to machine learning programs to make the case that the resources required to respond to these inductive challenges render critical aspects of their design constitutively value-laden. I demonstrate these points specifically in the case of recidivism algorithms, arguing that contemporary debates concerning fairness in criminal justice risk-assessment programs are best understood as iterations of traditional arguments from inductive risk and demarcation, and thereby establish the value-laden nature of automated decision-making programs. Finally, in light of these points, I address opportunities for relocating the value-free ideal in machine learning and the limitations that accompany them.

1 Introduction

According to a 2018 Pew Research Center survey, 40% of people believe that algorithmic decision-making can be objective, free from the biases that plague human decision-making.[1] Reading this result, one might reasonably ask what it really means for an algorithm to be objective and free from bias: what does it mean for an algorithm to be *value-free*? There are at least three interpretations of this question. On the least sophisticated interpretation,

[1] Smith 2018. 58% of respondents believe that computer programs will always reflect human biases.

we are asking whether algorithms operate wholly free of any influence of human values. The algorithms—we might answer—are just math; the data on which they operate are just facts; at no point in explaining their operation do we need to make reference to human values whatsoever. This, however, seems obviously false. Problematic social patterns unquestionably exist and are necessarily encoded in the data on which algorithms operate.[2] On a second and slightly more sophisticated interpretation, we might recognize the unavoidable encoding of such patterns in the data, and ask instead whether the algorithms themselves, their designs, are value-free. Those wanting to answer affirmatively might reason that, even if the data upon which the algorithms operate are shaped by human values, perhaps the engineers are still doing the best with what they are given by making value-free design decisions. On the other hand, those wanting to deny that design decisions are value free might do so by pointing to the all-too-human nature of the engineers themselves, arguing that engineers are inevitably subject to worldly pressures that in fact command the importation of human values, e.g., through the unavoidable capitalistic pressure to produce the highest profits for one's company. On this reasoning, algorithms could be value-free in principle, however—as a descriptive fact—they are not. There exists still a third interpretation of the question, on which we ask whether it is really possible for algorithms to be value-free *even in principle*. That is, is it possible for even a superhuman engineer—one impervious to worldly, self-interested pursuits—to produce an algorithm that is value-free? It is this third possibility that occupies me in this paper: when I ask if algorithms are value-free, I am asking whether values are constitutive of the very operation of algorithmic decision-making, such that on *no* idealized conception could they be value-free.[3]

Debates about at what point (if any) values can and should enter a decision-making procedure have been popular in various areas of philosophy. For example, the issue of to what extent epistemic norms in belief formation could be affected by practical and moral

[2] I discuss the relationship between these problematic social patterns and the operation of algorithmic bias in more detail in Johnson 2020a.

[3] Even this question will require some further precisification. In our academic musings, we can no doubt conjure up imagined algorithms operating on entirely fictitious data sets, whose decisions are totally divorced from real-world use. I'm not interested in asking whether those algorithms can be value-free. The set of algorithms that I am interested in includes paradigmatically algorithms that are used to replace or supplement human decision-making, that operate on real-world data, and whose decisions come to impact other human agents. Although possibly failing to include all conceivable algorithms, I still expect this class to be quite expansive, and certainly to include the algorithms that take center stage in discussions of algorithmic fairness and bias. I return to this point about the range of conceivable algorithms at the end of the paper.

aims has been widely discussed in literature on both pragmatic and moral encroachment.[4] In philosophy of science, prominent debates continue to unfold concerning whether values can shape not only the research programs scientists choose to pursue, but also practices internal to scientific inquiry itself, such as evidence gathering, theory confirmation, and scientific inference.[5] Ultimately this is a debate about whether values are a constitutive feature of scientific inductions. Like scientific inductions, machine learning programs use evidence (or known data) to form predictions (or generalizations to new phenomena).[6] Thus, there exists a natural but underexplored comparison between debates about objectivity in scientific inquiry and machine learning. In this paper, I take up this comparison by adopting arguments against the value-free ideal in science and extending them to the domain of machine learning. In so doing, I explore the extent to which machine learning algorithms are, can be, or should be value-free.

The literature concerning debates about values in science and the literature surveying philosophical perspectives on the use of machine learning programs are independently extremely vast. Thus, a comprehensive application of the former to the latter is well beyond the scope of any one paper. Instead, the aim of this paper is to build a bridge between two domains I anticipate will have much to contribute to one another. I begin this task by demonstrating how prominent arguments against the notion of scientific objectivity in the form of the problem of induction, underdetermination by evidence, arguments against demarcation, and the argument from inductive risk all straightforwardly apply to simple cases of machine learning use. My hope is that these comparisons will facilitate continued predictive and explanatory exchange between the algorithmic and scientific domains.

I begin in §2 by situating the discussion about values in algorithmic decision-making against the backdrop of the objectivity of induction more generally. In particular, I describe a notion of objectivity that all inductive procedures fail to meet due to the need to adopt bias in overcoming the problems of induction and underdetermination. It is against this backdrop that the so-called "value-free ideal" in science emerges. In §3, I present

[4] See Stanley 2005, Fantl and Mcgrath 2007, Moss 2018, Basu 2018, 2019, Bolinger 2018, Gardiner 2018, and Munton 2017, 2019a,b, among others.

[5] See Rudner 1953, Levi 1960, Douglas 2000, 2009, 2016, Rooney 1992, and Longino 1995, 1996, among others.

[6] Here and throughout I use 'machine learning programs', 'algorithmic decision-making', and 'algorithms' to pick out a broad class of automated programs that function by capitalizing on or "learning" from patterns manifest in the data on which they are trained in order to build a predictive model. This includes a wide range of machine learning programs, including supervised, unsupervised, and reinforcement learning programs.

two prominent arguments against the value-free ideal borrowed from feminist philosophy of science and argue for their application in the domain of machine learning programs. These arguments result in the view that both scientific and algorithmic decision procedures are deeply value-laden. In §4, I address a possible response on behalf of value-free ideal proponents: although many aspects of these decision procedures are value-laden, it might nonetheless be possible to relocate the value-free ideal by restricting it to just those processes responsible for updating probabilities in light of a fixed set of evidence. If so, then the value-free ideal is still arguably a desirable pursuit, albeit highly restricted. I end by addressing this reply, arguing that, although such a restricted role for the value-free ideal might be salvageable, its application to machine learning programs would render their use implausibly limited.

2 Origins of the Value-Free Ideal

In this section, I want to trace the historical progression of a pursuit of objectivity in scientific inquiry, and explore how it applies to the domain of machine learning. The unifying feature of the two domains is that both rely on induction. I regard as *inductive* any inference that is ampliative, i.e., that goes beyond the information given in the premises. This includes any inference that is non-deductive, and extends to both enumerative inductions, i.e., those that generalize from known instances to novel instances, and abductions, i.e., inferences to the best explanation. As we'll see, inductive inference is critical to both scientific theorizing and machine learning.

I start by discussing a notion of objectivity in induction that is undoubtably too strong, but seems surprisingly widespread in folk conceptions of scientific practice. This is a form of objectivity that Antony (2001, 2006, 2016)'s calls "Dragnet objectivity." The notion comes from the 1950s-1960s TV cop show *Dragnet*, in which Los Angeles Police Sgt. Joe Friday disciplines himself "to consider *just the facts*—the raw, undisputed data of the matter, unadorned with personal speculation and uncorrupted by emotional interest in the case," a strategy encapsulated in his famous catch-phrase "just the facts, Ma'am."[7] Applying this idea to scientific inference, the claim is that scientists should aim to favor hypotheses on the basis of "just the facts," without the influence of personal values. Applying this idea to machine learning, one can easily see why algorithmic decision-making is thought

[7] Antony 2006, 58.

to be more objective than human decision-making, since such programs are built to learn from raw data without the interference of personal speculation or emotional interest.

Although it makes for a quaint picture, Dragnet objectivity is an impossible model of scientific inquiry. The reason is that "raw data" by itself inevitably underdetermines various conclusions we might draw about some subject matter.[8] If ever we considered literally just the facts, we would never be able to draw inductive conclusions. Indeed, not even Sgt. Friday used Dragnet objectivity strictly speaking, since no human could. Induction is useful precisely because it is ampliative—it allows us to go beyond what is given to learn informative facts about the world.[9] Crucially, the evidence itself is always and in principle consistent with an indefinite (possibly infinite) number of different conclusions we could draw. No finite amount of data will ever be able to narrow the hypothesis space to one, since there will always be more than one hypothesis consistent with the data.[10] This is known as the problem of underdetermination of theory by evidence, and its roots can be traced back to another famous problem: Hume's Problem of Induction. To understand fully Hume's problem and the implications of it for the discussion of values, it helps to contrast properties of inductive and deductive arguments. Induction differs from deduction in two important respects. The first, which is often mistakenly taken to be the essence of Hume's problem, is that induction, unlike deduction, fails to guarantee truth. Because the conclusions of deductive arguments are always in some sense contained in their premises, if the premises are true, then the conclusion is guaranteed to be true. Inductions, on the other hand, are merely taken to provide probable support for some conclusion. If the premises of an inductive argument provide a confidence of 99% certainty in some conclusion, that still leaves 1% chance the conclusion could be false. So, in all inductions, there's a chance we could get things wrong.

However, this was not Hume's problem. Hume's problem concerns the second way in which induction differs from deduction: in justification. The justification of deduction is

[8] This is even granting the idea that data could be "raw" in some robust sense, presumably unadulterated by human collection practices. For a range of criticisms, see Gitelman 2013.

[9] The notion of Dragnet objectivity is intended to demonstrate an uncontroversial philosophical point: a list of facts alone will not allow us to form ampliative conclusions. It's uncontroversial because first, it's definitional on 'ampliative', and second, even the most stringent and traditional epistemological views will agree to it. For example, even objective Bayesians will agree that *something more* is needed beyond merely a fixed body of evidence, namely some procedure for updating in light of that evidence (for the Bayesian, Bayes' Rule).

[10] To put the point succinctly, for any theory T that is consistent with some body of observed evidence E, we can imagine another theory T' according to which E makes it seem like T is true, but it isn't. Evil demons and brains in vats are common resources for philosophers constructing such cases.

a priori and necessary. The justification of induction is not. Arguably, the justification of induction is contingent—it depends on the world being a certain way.[11] Thus, the problem with induction isn't that our degree of support always allows some room to be wrong, it's that there appears to be nothing to justify why known instances would provide support *to any degree whatsoever* to predictions of unknown instances. There's nothing *logically* at odds with the world suddenly becoming drastically different. Thus, whether some premises provide support for some conclusion depends on certain contingent features of the world, e.g., that the world continue to remain uniform and exhibit the patterns we've seen in the past and that are encoded in the premises. The problems of induction and of underdetermination generalize to any inductive procedure that attempts to use patterns present in evidence to make predictions about novel cases. Such inductive procedures are a critical aspect of scientific theorizing, but they're also a fundamental feature of machine learning. As we'll see, the feminist arguments in this paper stem ultimately from these related problems of induction and underdetermination.

Hume's problem of induction and the problem of underdetermination entail that no inductive process can be objective to the degree demanded by Dragnet objectivity. However, the success of science and induction more generally is a testament to our regularly overcoming both. Inevitably, we do this by taking certain facts for granted. By making non-evidential assumptions, the evidence can guide us toward some conclusions rather than others. The problems of induction and underdetermination apply to any domain of inquiry in which we attempt to draw conclusions on the basis of limited data, and each domain therefore comes with its own set of assumptions on which it relies. Examples of such assumptions include perceptual transformation principles (in the case of visual perception), cognitive heuristics, biases, and Bayes' Rule (in the case of belief formation), theoretical virtues, values, or paradigms (in the case of scientific inference), and parameters, filters, or constraints (in the case of machine learning programs), to name just a few. Elsewhere in my work, I call this broad collection of assumptions in different domains "biases".[12] However, to avoid debates about the normative connotations of the concept *bias*, going forward

[11] This is admittedly controversial. Norton (2003, 650, 666-669) makes a compelling case by adopting what he calls *the material theory of induction*, according to which "all inductions ultimately derive their licenses from facts pertinent to the matter of the induction", and demonstrating how it might evade the problem of induction. Of course, objective Bayesians, inductive rationalists, and logical probability theorists might all disagree. For notable critiques of these views, see the literature cited in footnote 34.

[12] Johnson 2020a,b. This notion of *bias* is intended to be normatively neutral (i.e., neither good nor bad necessarily).

I'll call this collection "canons of inductive inference."[13] What unifies them is that each serves as a "non-evidential way of limiting the hypothesis space to a tractable size."[14] In other words, canons of inference are necessary means of overcoming underdetermination.[15]

Induction requires the adoption of canons (in other words, "biases"); but crucially, canons are not one-size-fits-all phenomena. There are many possible bridges one might adopt to traverse the gap between evidence and theory, and there seem to be no *a priori* grounds for preferring some bridges over others. Famously, Hume's own response to the problem of induction was to claim that it was simply a "natural instinct" of humans to continue to perform inductions under the assumption that nature is uniform, so unobserved instances will share similarities with observed instances.[16] This canon has become known as "the principle of uniformity of nature," and it is widely adopted as one of the most fundamental bridges over the inductive gap. On a very blunt analysis of machine learning, this single assumption takes up the core explanatory burden for all cases of algorithmic bias. It is precisely the inductive assumption that the future will be like the past—in an environment whose past is shaped by historical injustice—that produces predictive results reflective of those historical injustices.

Of course, even saying that the future will be like the past is itself pretty unhelpful, since there are any number of ways in which one thing can be like another. Thus, the trouble with adopting the principle of uniformity of nature as an overarching solution to the problems of induction and underdetermination isn't that the principle is arguably false, but rather that it's trivially true, and does little by itself to reduce the hypothesis space to a tractable size.[17] So, the question remains of what other assumptions scientists need to adopt in order to accomplish the aims of science.

The debate within philosophy of science regarding the value-free ideal has centered around precisely this question. It's not a question of whether scientists must adopt *some*

[13] I borrow the notion of 'canons of inference' from Douglas (2016), who borrows it originally from Levi (1960). John Norton (2021, Ch. 5) adopts the notion of *criteria*, cautioning against the use of 'values' or 'virtues', since those connote that the canons could be ends in themselves (rather than means to a greater end, namely truth) and that the choice among them is entirely unconstrained and amounts to a free choice among scientific practitioners.

[14] Antony 2016, 161. Antony too uses the term 'bias' to pick out this general class, though she does not consider cases of machine-based decision-making.

[15] Cf. my functional account of bias, according to which, "what makes them all instances of the natural kind *bias* ... is that they all play the same functional role in overcoming underdetermination" (Johnson 2020b, 1216).

[16] Hume 1748, Section 5.

[17] This point serves as the basis of Goodman (1955)'s New Riddle of Induction.

assumptions—reasons we've discussed render that point undebatable. Rather, it's a question of *which* assumptions scientists ought adopt. In other words, a debate about which canons are acceptable and which are impermissible. A canonical answer to this question was provided by Thomas Kuhn. Kuhn (1962) famously argued that various scientific research programs come with their own package of assumptions on which to base scientific inferences, what he called a "scientific paradigm." Crucially, he additionally claimed that the only standards by which to evaluate these assumptions exist within the paradigms themselves. In other words, the question of whether some method is acceptable can only be answered from within a particular paradigm, and no one paradigm's answer to that question is better or worse than another's—the two are *incommensurable*. This radical claim eventually stuck Kuhn with the dreaded label of *relativist*.[18]

Wanting to avoid the relativist label, Kuhn (1977) eventually walked back his claims, and put forward a list of "theoretical virtues" he suggested could serve as paradigm-independent, objective grounds for theory choice. These include *accuracy*, *fruitfulness*, *consistency* (both internal and external), *breadth of scope*, and *simplicity*. These virtues have now come to be known as "epistemic values," and this haloed set was taken to provide at least a benchmark answer to the question of which canons scientists ought adopt: above all, they should adopt canons that promote epistemic values, which themselves are taken to promote truth. Adherence to this list is what historically gave rise to the so-called "value-free ideal" in science. The answer was benchmark since even Kuhn noticed that these values will themselves often trade off from one another. However, the critical insight that formed the basis of the value-free ideal was that in deciding what theory or hypothesis to adopt, scientists should always be guided by these epistemic virtues only, and not by social, ethical, or political values.

As Douglas (2016, 611) notes, the 'value-free ideal' starts to seem like a bit of a misnomer. She suggests a more apt label for the ideal would be 'epistemic-values-only-in-scientific-inference ideal'. There are two important points this new label highlights. First, as has been argued up to this point, there will always be some role for "values" (or canons (or biases)). However, those values (or canons (or biases)), according to the ideal, will be limited to the epistemic. The second important aspect of this alternative label is that it makes clear that the relevant focal-point of debates surrounding the value-free ideal is scientific *inference*. Everyone on both sides of the debate grants that values can guide

[18] See, for example, Lakatos (1970, 178)'s famous claim that Kuhn's theory rendered scientific theory choice "a matter of mob psychology."

some aspects of scientific practice, e.g., what research projects get taken up, what personal goals of individual scientists motivate pursuing those research programs (fame, fortune, etc.), or which demographics tend to populate which fields of research. Likewise, I take it that everyone will agree in the debate about whether algorithms are value-free that values shape some aspects of algorithmic use: what problems are addressed using machine learning programs, what overarching commercial aims individual companies have in producing algorithms, or which demographics tend toward or away from the technology industry more generally. These considerations fall "outside" of inductive inference itself, namely, the point at which we decide to accept or reject some conclusion, and thus, are irrelevant to the debate about the value-free ideal.[19] Thus, in what follows, I set them aside in order to maintain focus on what seems the best possible candidate for defending the value-free ideal, inference itself.[20]

It helps to take stock of the dialectic at this point. We began with a notion of objectivity that was a (surprisingly popular) caricature of science: Dragnet objectivity. According to this stereotype, science is a lot like deduction. When presented with the evidence that the world provides—just the facts—we immediately (or *a priori* and with certainty) know what conclusion is entailed by those facts. However, as we saw from the problems of induction and underdetermination, science is not like—and in principle cannot be—deduction.[21] Evidence alone (irrespective of canons of inference) is insufficient to establish ampliative conclusions. Thus emerged a more substantial and interesting target view of science: the value-free ideal. According to this view, we accept that both evidence and canons are essential to scientific inference, but we retain the value-free ideal of science by restricting the canons to just those values that are properly epistemic and preclude any canons that are shaped by social, ethical, or political values. Again, this is intended as a

[19] They are, of course, not irrelevant to the discussion about values in science and technology more generally. For a variety of perspectives on the relationship between values and these wider applications of technology in society, see canonical work by Friedman (1995, 1997), Johnson and Nissenbaum (1995), Friedman and Nissenbaum (1996), Nissenbaum (1996), Bowker and Starr (2000), Shmueli (2010), Kroll et al. (2017), Kroll (2018), and Mulligan et al. (2019).

[20] I will also be setting aside complications that emerge from the actual production of machine learning algorithms. For example, I set aside issues arising from inevitable performance errors and the fact that often algorithmic development is spread out across multiple, modular developmental teams, rarely with just one individual developer at the helm. The motivation here is the same: to identify the strongest possible case for the value-free ideal. I take it that if it can be demonstrated that in the best possible case the value-free ideal is untenable, then it will certainly be untenable in these more complicated scenarios as well.

[21] I don't mean that scientists never rely on deduction, only that induction is the canonical form of scientific inference.

theory of what is possible in principle. That as a matter of descriptive fact scientists often fall short of this aim won't do much to thwart the ideal. Thus, arguments against the value-free ideal in science strive to show that even in principle, this ideal is unattainable. There are two standard arguments that aim to establish this point, to be discussed in greater detail in the remainder of the paper. The first is the argument against demarcation: that the distinction between the epistemic and non-epistemic canons is untenable, because demarcating between the two is itself an essentially value-laden enterprise. As we'll see, this argument can be interpreted either as a claim about the justification of canons (that in justifying the demarcation, we're left with a further choice between epistemic and non-epistemic virtues, leading quickly to a regress) or as a claim about the constitutive natures of the canons themselves (that the canons are constitutively tied to value-laden features of the environment in which they're deployed). The second argument is from inductive risk: that any attempt to form conclusions on the basis of evidence will inevitably run the risk of getting things wrong, requiring an appeal to value-laden considerations for assessing this risk. If these arguments succeed, then like the caricature of Dragnet objectivity with which we started, the more substantial target view of the value-free ideal will need to be rejected.

The final critical insight of this section comes from recognizing that as inductive decision-making procedures, machine learning programs are subject to these same problems of induction and underdetermination. As the discussion of Kuhn brings out, like in the case of induction more generally, there will be no one solution to these problems in the domain of machine learning programs. To put the point bluntly, there can be no algorithm for building algorithms.[22] In fact, computer scientists are likely familiar with many of the points made here already, though perhaps not under the guise of Kuhnian theory.[23] Model

[22]This slogan, like most slogans, favors rhetoric at the expense of precision. It relies on an equivocation: I mean 'algorithm' in the first use of the word in the way Kuhn (1977, 359) means it when he says that an "algorithm able to dictate rational, unanimous [theory] choice" is "not quite [an] attainable ideal." The second use refers to the machine learning algorithms that are the target of this paper. The idea of the slogan is that you can (of course) write algorithms for building algorithms for something or other. What you can't do is write an algorithm for building algorithms for any arbitrary problem. And the problem of how machine learning algorithms should respond to data across the board—like the problem of determining what theory to adopt given some arbitrary set of evidence—is one of the latter problems.

[23]It's often thought that Hume's problem of induction resurfaces in the domain of machine learning in the form of the No Free Lunch Theorem (see, for example, Giraud-Carrier and Provost 2005, 2, Domingos 2012, 81, and Wolpert 2013, 2). I'm compelled by arguments made by Lauc (2019) that the NFL is more closely akin to Goodman (1955)'s New Riddle of Induction, though that too famously connected to Hume's problem in ways discussed. Either way, both interpretations lead to the result, argued for here, that in the domain of machine learning, "there is no learning without bias, there is no learning without knowledge."

builders undoubtably recognize that there are many different, yet acceptable ways to build predictive models.[24] There are of course norms within the professional community that go a large way toward restricting the domain of acceptable methods—these norms arguably comprise a Kuhnian paradigm. However, these norms cannot determinately settle every question, and some decisions about how these choice points play out ultimately depend on the goals of the predictive model together with individual aims of model builders. They might have to, say, choose between various types of regression models or make decisions about what cost functions to adopt. These can be interpreted as computer scientists having to make decisions about what canons to adopt. If program engineers adhere to the value-free ideal, then they're apt to produce programs that draw conclusions from some dataset in ways that maximize accuracy, fruitfulness, consistency, breadth of scope, and simplicity. The claim made at the beginning of this paper that algorithms are objective is, thus, charitably interpreted as the claim that so long as the decision points program engineers are responsible for are resolved in ways adhering to these canons, the algorithm itself is regarded as value-free.

The rest of this paper, then, explores the extent to which these decision points can be resolved by adhering only to the Kuhnian list of epistemic virtues, excluding entirely any appeals to social or ethical values. In what follows, I present objections to the value-free ideal in science. I then argue that these same objections apply to the adoption of the value-free ideal in the production, use, and evaluation of machine learning programs. Once the objections are in place, I explore how scientists have historically responded to these objections, and the extent to which these responses are available in the domain of machine learning.

3 Against the Value-Free Ideal

In this section, I present two famous arguments from feminist philosophy of science against the value-free ideal: the argument against demarcation and the argument from inductive risk. Following the presentation of each, I discuss how such arguments can be straightforwardly extended to machine learning algorithms.

(Lauc 2019, 484, echoing Domingos 2015, 64). For a greater discussion of these points, see Dotan 2020, where the NFL is used to motivated conclusions similar to those of this paper. Thanks to Kathleen Creel for helpful discussions about these points.

[24] The point applies to scientific modeling more generally. See, for example, Weisberg 2007.

3.1 The Argument Against Demarcation

The first argument I present against the value-free ideal is borrowed from the work of Helen Longino (1995, 1996) and disputes the very idea that we could demarcate between so-called "epistemic" and "non-epistemic" values in the first place.[25]

Longino (1995) makes this point by first curating an alternative list of social and ethical canons of inference drawn from historical work in feminist philosophy of science; this list includes values of *empirical adequacy*, *novelty*, *ontological heterogeneity*, *complexity of interaction*, *applicability to human needs*, and *diffusion of power*. She then goes through a stepwise comparison of values from the two lists, pitting traditional Kuhnian values and feminist social and political values against one another and demonstrating various contexts in which the feminist values ought be favored. For example, consider the feminist theoretical virtue of *novelty*, which Longino (1995, 385) argues requires new theories differ significantly from theories that are currently accepted. This requirement would seem to directly contrast with the Kuhnian virtue of *external consistency*, which requires new theories be consistent with theories that are currently accepted. Longino argues that both *novelty* and *consistency* can be viewed as appropriate virtues to shape theory choice; however, given that the two are at odds with one another, which virtue is adopted in any particular instance of scientific theorizing is a contextual matter, and crucially will be settled in virtue of the socio-political features of that context. As Longino (1995, 396) states, her arguments establish that "those traditional values are not purely epistemic (if at all), but that their use in certain contexts of scientific judgments imports significant socio-political values into those contexts."

Longino's arguments are complex, and there seem to me to be subtleties that admit of multiple interpretations. The most straightforward interpretation of the argument takes it as a point about socio-political values shaping the meta-decision about whether to choose values from one list over the other. In the case of *novelty*, the claim is that feminists adopt this virtue on the socio-political basis of aiming to depart from theories that have facilitated gender oppression throughout history. *External consistency*, on the other hand, is chosen on the socio-political basis of acceptance of the gender-oppressive status quo. Thus, in both cases socio-political values guide us (either wittingly or unwittingly) in accepting the canons that we do. No decision of which canons to adopt can ever be purely epistemic or value-free. This, she argues, renders a strict demarcation between the two

[25] A similar approach is taken up by Rooney (1992).

lists on the grounds that one set is value-free untenable, because the choice between the two lists is itself a value-laden decision. I regard this as *the justification argument against demarcation*: if your justification for choosing an epistemic virtue over a non-epistemic virtue (or vice versa) depends on social and political values, then a strict demarcation between the epistemic and the non-epistemic is untenable.

The second, more subtle interpretation of Longino's argument takes it as a point about the natures of the values themselves: that in virtue of being context-dependent, epistemic virtues necessarily imbibe the value-ladeness of the features of the environment to which they correspond in that context.[26] As another example, consider the contrast Longino (1995, 393-394) presents between *simplicity* and *ontological heterogeneity*. Using a more recent example, we can see this tension arise in cases of clinical drug trials. Consider the case of Ambien. In 1992, the prescription drug Ambien was approved as a sleep aid by the FDA. However, clinical trials for Ambien didn't take into account the average metabolic differences between men and women, resulting in a recommended dosage that was the same for both. Twenty years later, research on the effects of sleep aids on impaired driving found that women were being prescribed nearly twice the amount they should be, resulting in many women who took Ambien at night to still have enough of the drug left in their systems the next morning to impair their abilities to operate a motor vehicle.[27]

Because researchers took male metabolic systems to be the paradigmatic case, an assumption driven by an allegiance to canons that maximized the epistemic value of simplicity, they produced a drug that put the lives of many women taking the drug in danger, a mistake that took two decades to rectify. Crucially for Longino, scientists' decision to posit the fewest kinds of entities in a context where members of a privileged class—in this case, males—are taken to be the primary fundamental entity, lead to theories that both

[26] This interpretation is a potential departure from Longino's original intentions for the argument against demarcation. Minimally, we can say it takes as inspiration Longino's original argument. I personally find it the most compelling case against the value-free ideal. This interpretation likewise shares important similarities in structure to Norton (2003, 2021)'s *Material Theory of Induction*; though he would undoubtably object to its use in drawing conclusions against the VFI. In future work, I hope to explore this interpretation's relationship to externalist theories of empirical warrant more generally (see, for example, Graham and Pedersen 2020, particularly Burge 2020 in that volume).

[27] Tavernise 2013. A similar story plays out in the differential efficacy of HPV vaccine Gardasil in white women and Black women. Gardasil primarily targets HPV types 16 and 18, while Gardasil 9 targets types 16, 18, 31, 33, 45, 52, and 58. However, Black women are about half as likely to be afflicted with types 16 and 18, and neither form of vaccine is effective against three types of HPV that most commonly afflict Black women, 35, 66, and 68 (Vidal et al. 2014). This is arguably one contributing factor to rates of HPV-associated cervical cancers being higher among Black women than among white women (Viens et al. 2016, 662).

legitimate and perpetuate the socio-political values on which they're built. In a world built on socio-political values that result in a hierarchy where some members of the population belong to a privileged class, those same individuals get prioritized in theory, and thus an allegiance to simplicity will (either wittingly or unwittingly) imbibe the very socio-political values on which the hierarchal relations are formed. Thus, to demarcate the non-epistemic from the epistemic in any particular context is untenable. I regard this as *the constitutive argument against demarcation*: if the adoption of any seemingly epistemic virtue in a particular context depends constitutively on the socio-political features of the context, then a strict demarcation between the epistemic and non-epistemic is untenable.[28]

Although never explicitly discussed by Longino, I take the demarcation argument (in both forms) to be deeply related to Hume's problem of induction, and in particular to the second aspect of how induction differs from deduction: induction, unlike deduction, is justified contingent on how the world is. Focusing first on the constitutive argument against demarcation, there's no such thing as an induction that's justified *a priori* or irrespective of facts about the world. Rather, whether any particular induction is justified seems to depend on features of the world. In this way, features of an induction, e.g., being justified, seem inseparable from features of the world, e.g., being uniform. There's no such thing as justification in the abstract. There's only justification that depends constitutively on the world's being a certain way. So too, I like to think of Longino as claiming other features we take to be independent features of inductions, e.g., being simple or being externally consistent, are inseparable from features of the world, e.g., promoting the gender oppressive status quo.[29] If induction is justified only by contingent features about the world, and the

[28] I adopt "constitutive" in order to highlight how externally-grounded empirical warrant is constitutive of inductive inference's ability to conduce to knowledge (Burge 2020, 39). By its use, I don't mean to suggest that non-epistemic values constitute epistemic values, but rather that non-epistemic values are partly constitutive of the warrant given by any inductive inference, even one apparently based solely on epistemic values. To clarify further, while it's true that an epistemic value's relationship to any particular non-epistemic value is contingent, its relationship to non-epistemic values generally is constitutive in this way. For example, had our society favored women as the paradigmatic human kind, simplicity would then perpetuate a matriarchal value structure. But, while the favoring of any particular group is contingent, the favoring of some group or other, or the adoption of an egalitarian sample, is necessary. So it is in this sense that I take the overlap between some epistemic value (e.g., simplicity) and non-epistemic values (i.e., patriarchy, matriarchy, or egalitarianism) to be inevitable and, thus, "constitutive".

[29] To make this point in another way using a framework familiar to computer scientists, consider the Proxy Problem: often seemingly innocuous attributes that correlate with socially sensitive attributes, serve as proxies for the socially-sensitive attributes themselves. (I discuss this problem at length in Johnson 2020a, see also Hellman 2018, 2020b and Hu and Kohler-Hausmann 2020.) We could think of properties of inductive arguments as necessarily serving as proxies for socio-political features of the world. Not only do properties of objects unwittingly encode socially-sensitive properties, but also properties of inductions can too. So,

world is itself shaped by our values, then our values necessarily influence which canons will be most likely to generate justified inductive arguments. Crucially, which values are apt, even narrowly for the epistemic aim of getting us onto truth, will itself be a contingent matter; it depends ultimately on how the world we live in happens to be.[30] Given that how simple (or externally consistent) some induction is will be a relation between the theory and the world, the fact that the world has been shaped by various forms of oppression entails that adherence to these epistemic values will de facto result in adherence to those social and political values.

Likewise for the similarities between Hume's problem and the justification argument against demarcation. As mentioned in the previous section, Hume's favored response to the problem concerned the brute adoption of the assumption that nature will continue to be uniform, i.e., the Principle of Uniformity of Nature.[31] Now let us ask: on what grounds might we justify the adoption of that assumption? An obvious response would be that such a premise has held true for us in the past, and so we can expect that it will continue to hold true in the future. But this is to use induction to justify induction, which would beg the question. One could think of Longino as making a similar point: proponents of the value-free ideal neglect to recognize how their choice to adopt some canons of inference over others is itself a value-laden judgement, one that comes equipped with various biases in the form of allegiances to accuracy, simplicity, breadth, etc. As Longino's argument suggests, there might be contexts in which such biases are apt. However, there might likewise be contexts in which biases in the form of other canons, such as ontological heterogeneity and applicability to human needs might be more apt. Any attempt to justify adherence to only epistemic virtues in the meta-decision about which list to adhere to would itself require justification or else be question-begging. This leads to an impossible regress in

just as some property about objects we're interested in (zipcode, say) might unintentionally be a proxy for some other feature of that object (race), so too, I contend, some property about an induction (aiming for simplicity) might unintentionally be a proxy for some other feature of that induction (perpetuating patterns of injustice). So something like simplicity might appear value-free, but when you consider it against the backdrop of the subject matter in which a decision procedure is adopted, it too becomes an unwitting proxy for other properties, like the perpetuation of patriarchal structures.

[30] This point is related to the "tradeoff" other theorists have argued exists between *epistemic reliability* and *justice*. If our standards for accuracy have been shaped by the social environment, which has itself been shaped by oppressive structures, then there will arguably exist a tradeoff between demands of epistemic reliability and demands of morality. See Gendler 2011, Basu 2018, 2019, Bolinger 2018, Munton 2019a, and Johnson 2020a, among others.

[31] This is bracketing Goodman (1955)'s point that the principle is trivially true and, thus, ultimately unhelpful in resolving the problem of induction.

justification.[32]

Crucially, the arguments against demarcation seem to apply equally well in the case of machine learning. As mentioned in the previous section, machine learning engineers will have certain decision points left up to them. Ultimately, the decision to use some data analysis method over others will depend on the aims of the program and the goals of the programmer. Here, the question of how to justify some methods over others will likely be answered by appeal to the value-free ideal: the decisions to, say, use a parametric rather than a non-parametric model might be guided by the fact that the former is simpler than the latter.[33] Alternatively, one might adopt a non-parametric model due to its flexibility, opting for a predictive model that more closely aligns with ontological heterogeneity. Crucially, the rationale for choosing one over the other seems itself open to scrutiny and calls for justification. It is in providing this further justification that program engineers will likely have to appeal to facts that go beyond the purely epistemic.[34] They often include considerations about the overall aim of the program and the context in which it is intended to be used, facts which themselves depend on social and political factors. According to the justification argument against demarcation, any further justification that involves social or ethical considerations will render even those first-order decisions value-laden in significant ways. Moreover, from the constitutive argument against demarcation, even abiding by a seemingly pure epistemic list of considerations when making design decisions might usher in socio-political values. A straightforward example of this is in the selection of a loss function: this selection ideally corresponds to actual expected loss in making an incorrect prediction; in this way, we want the function to accurately approximate loss. However,

[32] I take it the nail in the coffin for the value-free ideal on this interpretation of the argument would be to demonstrate that non-epistemic values alone can end the regress. I'm told that this sort of argument might be constructed by drawing on Plato's argument for the priority of the Good. I'm not familiar enough with that argument to make the case here. However, a more flat-footed approach might draw on Hume's own response to the problem of induction to point out that justification has got to stop somewhere (see also Ward 2021, 60); eventually we have to proceed with scientific theorizing. Surely the decision to cut off justification at any particular point will therefore be a pragmatic decision, and thus one that depends on non-epistemic values. Thanks to Jim Kreines for helpful discussion of these points.

[33] Of course, what *simplicity* ultimately amounts to is itself an elusive question in the history of philosophy of science.

[34] This point can be bolstered by literature on probabilistic reasoning more generally that converges on a similar point. A small sampling of that literature includes Ramsey (1989)'s criticism of Keynes that there are no objective probabilities, Carnap's failure in constructing a purely formal foundation of inductive logic (for summary, see Zabell 2011), Titelbaum (2010)'s rejection of a purely objective notion of evidential support, and Fallis and Lewis (2016)'s critique of purely objective measures of probabilistic accuracy, among many others. Thanks to Branden Fitelson for pointing me to this literature.

there are many dimensions on which to measure actual loss, corresponding inevitably to value-laden features of the world. Adoption of any particular loss function so as to approximate real-world loss will, according to the constitutive argument, necessarily imbibe those loss functions with the very socio-political values of the real-world losses.

Whereas I've described the demarcation argument(s) as related to Hume's problem of induction and the fact that inductions (unlike deductions) are contingent, justified only relative to the world's being a certain way, the next argument from feminist philosophy of science against the value-free ideal focuses on the other way induction differs from deduction: in the potential for getting things wrong. According to this argument, any inductive reasoning *must always* culminate in a decision that involves ethical considerations. This is because at some point in the chain of justification, one must consider the possibility that the prediction being made might get things wrong, a risk that inevitably comes with social and ethical costs. It is the consideration of this risk and the consequences these arguments have for the value-free ideal that I turn to next.

3.2 The Argument from Inductive Risk

As discussed in §2, canons of inference arise from a need to overcome underdetermination; they are bridges over inductive gaps. Importantly, however, such bridges are inevitably fallible. This is in the nature of biases and heuristics in general: they give us guidance in cases where we're unsure, i.e., cases where we're not guaranteed truth.[35] Such canons can never guarantee truth, since to do so would be to turn induction into deduction, and that would be tantamount to omniscient knowledge about the way the world will be in unobserved cases. So, in all cases where we adopt canons of inference, i.e., in all cases of induction, we run the risk of getting things wrong—we call this "inductive risk," and it prompts the second argument against the value-free ideal.

The argument from inductive risk can be traced back to Richard Rudner, who states:

> [S]ince no scientific hypothesis is ever completely verified, in accepting a hypothesis the scientist must make the decision that the evidence is *sufficiently* strong or that the probability is *sufficiently* high to warrant the acceptance of the hypothesis. Obviously our decision regarding the evidence and respecting how strong is "strong enough", is going to be a function of the *importance*, in the typically ethical sense, of making a mistake in accepting or rejecting the

[35] Cf. Johnson 2020b, 1217, footnote 44.

> hypothesis. ... *How sure we need to be before we accept a hypothesis will depend on how serious a mistake would be.*[36]

The idea here is that, contra the value-free ideal, ethical values have a legitimate and necessary role to play in guiding scientific inference, because they establish confidence thresholds for ultimately accepting or rejecting a given hypothesis or prediction. This point is made obvious by comparing two hypothetical scenarios. Imagine that in one case, engineers are responsible for producing seat belt buckles, while in another scenario, engineers are responsible for producing pant belt buckles.[37] In both cases the engineers run the risk of getting things wrong in producing defective buckle designs, but the consequences of getting things wrong in the former case are much more dire than those of the latter. Clearly the threshold for confirmation in the two cases should not be the same: we should demand a much higher degree of confidence in the engineers' hypotheses in the first scenario than in the second. Proponents of the argument from inductive risk insist that the threshold for confidence can only be established by appeal to ethical values, thus rendering the decision to adopt any particular hypothesis value-laden.

This line was picked up by feminist philosophers of science, most notably in the work of Heather Douglas (2000, 2003, 2009, 2016). Douglas expands on Rudner's initial point by arguing that not only do scientific inferences involve such risk, but that the risk is compounded by the weight given to scientific judgements as expert testimony in various social and political arenas. As Douglas (2016, 615) states, the "baseline epistemic authority brings with it general responsibilities to be neither reckless nor negligent in one's actions ... scientists must consider in particular the impact of their authoritative statements."[38]

In sum, not only do scientists have to take into consideration the risk of getting things wrong whenever they perform an inference to some hypothesis, but also they have to keep in mind the influence that that wrong hypothesis will have in communities in which their judgement is regarded as expertise. Importantly, all these points apply equally well, if not more so, in the case of machine learning programs. Consider a minimal pair similar to Rudner's original case, but now in the domain of machine learning. For example, imagine

[36] Rudner 1953, 2, emphasis in original.

[37] This is an adaption of Rudner (1953, 2)'s original case.

[38] See also Douglas 2003, 2009, chapter 4. Arguments from inductive risk in philosophy of science share important structural similarities with arguments for moral or pragmatic encroachment in epistemology (Stanley 2005, Fantl and Mcgrath 2007, Moss 2018, Basu 2018, 2019, Bolinger 2018, Gardiner 2018, and Munton 2017, 2019a,b, among others). One could thus use this section as a model for how to likewise extend a theory of moral/pragmatic encroachment to algorithmic decision-making.

you're tasked with building an image recognition program to distinguish human shapes from non-human shapes. The level of inaccuracy you should tolerate will depend on the use to which your algorithm will be put. If it is to be implemented in an office complex as a trigger to activate automated lights, 70% accuracy would be inconvenient, but acceptable. However, if it is to be implemented in a self-driving car to prevent pedestrian collisions, you should demand near perfection. Algorithmic design decisions about how to manage error therefore inherently involve values. We also see the analogue of Douglas's point in this domain: not only do machine learning programs run the risk of getting things wrong (a risk the negative consequences of which have been well-documented within the machine learning activist communities), but also because of the computational prowess, efficiency, and ubiquity of machine learning programs, we can expect the effects of their judgements to be wide-reaching and vast, likely more so than any individual scientist's judgements.[39] Thus, in building machine learning programs, it seems it is not sufficient that a program engineer merely adopt an aim of achieving some traditional canon like accuracy, since there's no such thing as accuracy *neat*. Instead, how accurate is accurate enough necessarily involves some determination of the ethical consequences of getting things wrong.

3.3 Application

Before moving onto possible responses to these arguments against the value-free ideal, it helps to demonstrate the application of these points to a concrete case in machine learning. One particular case that is popular in discussions of algorithmic bias and fairness concerns criminal justice risk-assessment algorithms such as COMPAS. COMPAS, developed and owned by Northpointe (now "Equivant"), stands for the Correctional Offender Management Profiling for Alternative Sanctions. It is a program used by judges across the United States to produce recidivism risk scores. The program works by collecting data about defendants awaiting trial and, on the basis of statistical analysis, produces a risk score that can then be used by judges to make decisions about, among other things, setting bail, establishing the need for pretrial detention, sentencing, or parole. However, a 2016 exposé by ProPublica revealed that in an analysis of over seven thousand COMPAS uses, the program was almost twice as likely to falsely label Black defendants as future criminals

[39] For a broad discussion of the impacts of algorithmic decision making on discrimination, law, and policy, see Barocas and Selbst 2016, Kroll et al. 2017, Selbst et al. 2019, and Abebe et al. 2020, as well as other scholarship produced in association with the Fairness, Accountability, and Transparency in Machine Learning (FAT* ML) community.

than white defendants, while often mislabeling white defendants as low risk at a higher rate than Black defendants.[40] This prompted many advocates for fairness and transparency to regard the program as problematically biased against Black Americans.

However, further analysis proved that this initial assessment was not so simple. Computer scientists researching the so-called "impossibility result" note that there are (at least) three candidate notions of fairness between two groups of interest one could adopt in the production of a risk-assessment program: first, its being *well-calibrated*, i.e., it has the same degree of accuracy within each group; second, its having *false-positive equality*, i.e., the same proportion of false positives within each group; and third, its having *false-negative equality*, i.e., the same proportion of false negatives within each group.[41] The impossibility result is that these three conditions cannot all be satisfied at once; they necessarily trade off from one another. Since this discovery, a heated debate has emerged regarding how best to balance these potential notions of fairness and, given that we can't have all three, which among the identified criteria of fairness—accuracy, balancing false positives, or balancing false negatives—should be given the most weight.

An extended philosophical discussion of the debates surrounding COMPAS and other criminal justice risk-assessment software is unfortunately beyond the scope of this paper.[42]. Instead, I intend only to apply lessons from the above discussion to demonstrate that two points can be made to gain traction on this debate. The first is that our preliminary discussion about the role of canons in inference establish that it would be a mistake to expect a monolithic set of conditions that all risk-assessment algorithms ought satisfy. To repeat an important point, there can be no algorithm for building algorithms. This contention is bolstered by the fact that the three intuitive conditions of fairness identified are at the same time reasonable and incapable of being jointly satisfied. Moreover, as we've seen from the argument from demarcation, reasonable canons of inference often pull in opposite directions, and whether we opt for some over others will ultimately be a contextual matter.[43]

[40] Angwin et al. 2016. This analysis compared risk scores produced by COMPAS with occurrences of rearrest over the following two years.

[41] This "impossibility result" has been formally demonstrated in the work of both Chouldechova (2016) and Kleinberg et al. (2016). Typically the relevant groups will be two that differ from one another along some salient social dimension, e.g., race. Kleinberg et al. label the candidate notions of fairness 'well-calibrated', 'balance the positive class', and 'balance the negative class', respectively.

[42] For detailed discussion, see Hellman 2020a

[43] A cottage industry has since developed within machine learning literature identifying and exploring novel notions of fairness beyond Chouldechova (2016) and Kleinberg et al. (2016)'s initial lists. Mehrabi

The second point concerns which criterion we ought to adopt in the case of COMPAS and the context in which it is used, namely the criminal justice system. Notice that those who advocate for giving greater weight to balancing false positives are in essence reiterating the traditional argument from inductive risk. Proponents of this view claim that the risk for the Black community of making a mistake and falsely flagging Black defendants as likely to reoffend is high enough that we ought to establish a greater threshold of confidence in making these assessments (or refrain from making them at all). Moreover, it's important to note the point made by Castro (2019, 417-418) that, in the current context of use, namely, the United States criminal justice system, "it's already disproportionately costly to be black." Thus, it's critical that a full assessment of inductive risk take into account not only the harms of getting it wrong that are shared among Black and white defendants, but also the differential harms on the Black community and Black defendants specifically when they're misidentified as high-risk. Moreover, defenders of the use of criminal justice risk-assessment tools like COMPAS likewise use arguments about risk. On this side of the issue, experts argue that to ignore the results of COMPAS and to instead rely on the judgements made by individual judges (unaided by risk-assessment tools) would likely result in decisions that are *more* problematically biased, not less.[44]

Again, it is impossible to settle these disputes here. Instead, I use this discussion to make the more modest point that value-laden debates are already well underway in the domain of machine learning. That both sides attempt to use considerations of social fairness and ethical risk in determinations of which notions of fairness to adopt, bolsters the claim that decisions about what canons of inference to adopt in machine learning—or even about whether to deploy the program for use in the context of the criminal justice system at all—will necessarily involve value-laden considerations.[45]

et al. (2020, 11-12), for example, count at least ten different notions common to the literature. This trend echos that in philosophy of science literature of identify and explore novel notions of theoretical virtues beyond Kuhn (1977)'s initial list. Keas (2018, 2762-2763), for example, counts at least twelve different notions common to the literature. Thanks to Katie Elliott for drawing my attention to these parallel trends.

[44]See, for example, Flores et al. 2016, 17 and Corbett-Davies et al. 2017. It's important to note also that the aforementioned impossibility result is a mathematical generalization that applies to any decision-making procedure, whether human- or machine-based. So, the same sorts of disputes ought reoccur in analyses of decisions made solely by human judges.

[45]Another interesting project emerges from considerations of adopting traditional feminist theoretical virtues in the domain of machine learning. For example, looking back at Longino's original list, we see as one possible canon of inference the value of applicability to human needs. Longino (1995, 389) describes this criterion as a call for scientific inquiry directed at reducing ongoing harms and "improving the material conditions of human life, or alleviating some of its misery." She gives examples such as pursuing scientific

4 Relocating the Value-Free Ideal

A popular defense of the value-free ideal in science begins by conceding part of the argument from inductive risk.[46] It accepts the point that the act of accepting or rejecting a hypotheses necessarily incurs inductive risk and, therefore, requires appeal to ethical values. However, it is for this reason, defenders argue, that scientists should never aim to accept or reject hypotheses at all, but should instead only ever assign probabilities to hypotheses with respect to a fixed set of evidence.[47] Following this line of thought, one might argue that in the case of COMPAS, the program's use is only ever to give a confidence rating that a defendant might recidivate, but it should not aim to make a determinate prediction one way or another. In fact, this arguably accords more closely with the stated use of the program. For example, in its Practitioner's Guide, Northpointe (2015, 7) introduces the AIPIE model of procedures for case management involving assessment tools like COMPAS. The AIPIE model has five steps: **A**ssessment, **I**nterpretation of results, **P**lan based on the information gathered, **I**mplement the plan, and **E**valuate the results of the action. This model utilizes tools like COMPAS at the first step—in Assessment—suggesting that the remaining steps of case management are left up to practitioners. Thus, a defender of the value-free ideal might argue that the considerations of the risk of getting things wrong come only in the interpretation, planning, and implementation phases. These are the steps wherein judges make predictions about whether particular defendants will reoffend and, although they do this by interpreting the results provided by COMPAS's risk scoring system, the risks associated with the predictions are not inherent to the algorithm itself. One way to see this is to notice that there can be several different decisions a judge might rely on COMPAS risk scores to make. They might use COMPAS risk scores in deciding who should be granted bail, or alternatively they might use it to decide to whom they should allocate

ends that produce sustainable agriculture, reverse the destruction of the environment, and assist the infirm. In the context of COMPAS, one can use this value to make the case that software like COMPAS—software that falsely identifies Black defendants as high risk, compounds the harm done to the Black community, and perpetuates historical patterns of oppression and injustice—is antithetical to such aims. Thus, the specific appeal to feminist theoretical virtues provides a platform for condemning reliance on this software in the context of the criminal justice system.

[46] The line of defense that I'm about to recap here focuses narrowly on the argument from inductive risk and leaves unaddressed the arguments against demarcation. I'll return to this point at the end of the section.

[47] Jeffrey 1956. Putting the point in familiar Bayesian terms, one might argue that the role of a scientist is not to accept some hypothesis H, but merely to assign some conditional probability for H relative to some body of evidence E, standardly in accordance with Bayes' theorem: $P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H)+P(E|\neg H)P(\neg H)}$

resources intended to prevent recidivism.[48] This suggests that inductive risks associated with the implementation of risk scores are associated with the decisions that are made by judges using the program as evidence, and not risks inherent in the assessment produced by the algorithm alone. Decisions about how and to what extent the risk scores are used in further decisions about the defendant are beyond the control and, thus, ethical purview of the program's architects.

This seems like a natural response to the argument from inductive risk. However, it's so natural that Rudner anticipates it in his original article. There, Rudner (1953, 4) preemptively gives the response that even in assigning some probability p, the scientist is accepting a kind of hypothesis, namely, a hypothesis "that the degree of confidence is p."[49] He claims this new hypothesis itself comes with a risk of getting the confidence wrong and, thus, the argument from inductive risk iterates. The move to confidences or probabilities only pushes the problem back a level, it does not escape it. Looking back on Douglas's arguments that scientists are taken as epistemic authorities, we might look to evidence of to what extent mere probability estimates in fact determine the judgements made on the basis of a scientists' testimony, regardless of whether it was the intent of the scientists themselves. This sort of default authority is clearly evident in cases of criminal justice risk-assessment scores like COMPAS, where analyses indicate that the products of algorithmic risk assessment programs do in fact sway judges' rulings, often resulting in greater sentencing disparities among individuals from marginalized groups.[50] Thus, the assignment of risk scores will itself come with various inductive risks when the algorithm is taken as an epistemic authority. Even if we interpret the assignment of risk as a mere assignment of confidence in, or probability of, recidivism, that assignment will still arguably entail inductive risk.

However, defenders of the value-free ideal in science push back: some argue that it is a conceptual mistake to regard the assignment of probabilities as the adoption of any sort of hypothesis.[51] The assignment of probabilities doesn't involve any *act* of acceptance, but rather merely *reflects* the degree of confidence we can mathematically hold in some prediction being true. Thus, there's no act to assign risk to, and the mere reflection of confidence is free from the argument from inductive risk. In the context of actual

[48] As one anonymous referee points out, feminist considerations such as those presented in footnote 45 might be naturally subsumed under this aspect of COMPAS's intended use.

[49] See also Douglas 2009, 53-54 and Douglas 2016, 610.

[50] See, for example, Skeem et al. 2020. See also Logg et al. 2019 and Christin 2017 for relevant discussion.

[51] Jeffrey 1956, 246. See Steel 2015, 82-83 for discussion.

scientific practice, these assignments of confidences can be seen as reflecting the actual confidences scientists have in particular hypotheses, and not an act or decision made on behalf of the scientist. Opponents of the value-free ideal in turn have a response: they claim that this interpretation of confidences requires conceptualizing scientists as idealized Bayesian agents, when in fact they are not. Bayesian models are poor models of the actual confidences scientists hold, and so in practice scientists do end up doing something akin to accepting a hypothesis, even when they assign probability estimates.[52] However, one might reasonably argue that this is a point of departure between the domain of scientific inquiry and machine learning. Although it's true that individual scientists fall far short of idealized Bayesian agents, given the computational prowess of machine learning programs, perhaps their predictive models can more closely approximate such agents. This would suggest that their assignment of confidences might well be mere reflections of mathematical confidences, reflections that do not admit of inductive risk.[53]

Although this is an interesting line of argument, ultimately investigating to what extent algorithms can instantiate idealized Bayesian agents without ultimately relying on values creeping into the models is beyond the scope of this paper.[54] I suggest it merely as one path defenders of the value-free ideal in machine learning might pursue. I instead end by considering what happens when we grant this point, accepting that we might relocate the value-free ideal by carving out a minimal role for it in the mere assignment of probabilities.

I argue that, even granting this point for the sake of argument, this still won't render most machine learning programs value-free. This is because most of these programs are used to automate decisions, not merely in assigning probabilities, but in making predictions about new cases and applying those predictions in the classification of novel stimuli. Proponents who adopt this as a potential role for the value-free ideal in machine learning *and* maintain that algorithms should remain objective and value-free will not only greatly limit the potential role for algorithms in automated decision-making, but will be forced to accept that the vast majority of machine learning programs in use today already outstrip the resources afforded to strictly value-free processes. Thus, by these standards, not many

[52] Steel 2015, 85-87.

[53] Although see Parker and Winsberg (2018) for an argument that it is precisely due to this computational prowess that decisions made earlier in the model-building process quickly compound and exacerbate risks of error, risks that are eventually inherited in the assignment of confidence measurements. Similarly, Steel (2015, 85) argues that any complex Bayesian statistical analysis will necessarily involve decisions among types of probability distribution, and that it is in these decisions that arguments from inductive risk iterate.

[54] Cf. footnote 34 for relevant literature in formal decision theory.

algorithms can be value-free, since their very use and function depends on their going beyond what a purely value-free process could accomplish. Algorithms don't merely create probabilistic estimates, they *use* those estimates to make classifications and produce outputs that make predictions about novel cases. This coupled with the recognition that such a line of defense for the value-free ideal leaves wholly unaddressed the arguments against demarcation makes the case for the ideal even more dire. If those arguments are correct, then arguably the design methods that result in the relevant probability assignments will themselves be either justified by or constitutively dependent on value-laden features of the world.

5 Ethical Reflections

Where does all this leave us, *morally speaking*? Admittedly, my goal has not been to advance any distinctively normative claims, but rather to demonstrate, through the application of feminist insights in philosophy of science to machine learning research, the much-needed role of moral philosophers in these debates about algorithmic decision-making. However, some resulting normative prescriptions are clear. Most obviously, if the arguments presented in this paper succeed, then algorithms cannot in principle be value-free. Granting that ought implies can, this entails that it's not the case that algorithms *ought* be value free. By foreclosing the normative question of whether algorithms should be value-free, we are now in a position to usher in the moral evaluation of algorithms.

This endeavor of exploring what normative conclusions can be drawn works best by case study. Consider, for example, a concrete attempt made by Johndrow and Lum (2019) to eliminate unfairness in recidivism risk programs. Their method for eliminating bias relies on creating statistical parity among the relevant classes, in the context of the recidivism prediction tool that they analyze, this would mean creating statistical parity among white and Black defendants, producing what they call race-independent recidivism prediction. What fairness amounts to is "differences in the distribution of the model's predictions conditional on the protected variable do not exceed some pre-determined threshold."[55] Ultimately, the relevant threshold the authors adopt is zero, which entails statistical parity between the two classes. However, they note that the decision to focus on statistical parity is motivated by pre-theoretical considerations about the inability to obtain data about

[55] Johndrow and Lum 2019, 214.

actual crime rates, instead needing to rely on imperfect proxies like rearrest, which are known to encode systemic biases against Black people. Thus, the decision to use statistical parity in the context of criminal justice is a considered one, which they acknowledge is open to scrutiny, and that is ultimately up to "policymakers and ethicists" to evaluate.[56]

Regarding import from the discussion of values in this paper, we can see how this decision to leave some vital parameter of the algorithm to be decided by ethical consideration is licensed by the conclusion of the arguments presented above that algorithms are inherently value-laden. With little imagination, one can envision an individual who subscribes to the alternative idea that algorithms can and should be value-free (perhaps one among the 40% of the population referenced in the first line of this paper) objecting to Johndrow and Lum's approach on principle. By building in a placeholder for ethical demands to make a contribution to how the algorithm operates, they might argue, the algorithmic design violates the normative prescriptions of the value-free ideal. By demonstrating that this ideal is in principle unattainable, the conclusions of this paper forestall this critique straightforwardly, and sanction the contributions of explicitly normatively-laden facets of the algorithm itself. Thus, the general approach of deliberately incorporating explicit value-laden choice points is vindicated.

However, other conclusions of this paper suggest limitations to other aspects of their general approach. In particular, the basic premise of their strategy can be viewed as an attempt to equally distribute inductive risk by operationalizing on and eliminating disparate outcomes for white and Black populations. However, it's not obvious that the best way to mitigate inductive risk is to adopt a policy of distributing equally across populations. Considering Longino's demarcation arguments, one might worry that this approach taken as a general strategy fails to be sensitive to how the question of which values one ought adopt will ultimately be a contextual matter.[57] One major and recurring lessons of the paper is that we should stop expecting a universal set of standards that dictate which values algorithmic decision procedures ought maximize across the board.

Indeed, this is precisely what we see when we reflect on how to expand a commitment to statistical parity as a strategy for mitigating risks across multiple groups. Inevitably, such attempts will likely trade off from one another, even localizing to a particular context like criminal justice. Johndrow and Lum (2019)'s algorithm creates "fairness" by ensuring statistical parity with respect to race in the outputs. However, even granting that

[56] Johndrow and Lum 2019, 194, fn. 3.

[57] To their credit, Johndrow and Lum (2019, 191-194, 214) continually note this limitation of their model.

statistical parity is what we want in the domain of race, this same general strategy does not straightforwardly apply to other social groups.[58] For example, given that overall rates of criminality are smaller for women than men, we might expect that a commitment to statistical parity between male and female defendants is misguided.[59]

Once we recognize that differences between social groups matter for the application of the model, we can anticipate that these complications are only compounded when we consider individual defendants' occupying intersectional identities. These different identities might very well pull in different directions with respect to the need to mitigate inductive risk, and it's not clear the intersection of different identities admit of any easy quantitative analysis. Thus, it's difficult to predict what effects Johndrow and Lum (2019)'s allegiance to erasing indicators for race will ultimately have on individuals who occupy multiple marginalized groups, some of which should afford them more or less favorable treatment when calculating inductive risk. More importantly, for these reasons it seems that calculations of inductive risk more generally don't admit to straightforward operationalization of any kind, again bolstering the need for experts well-versed in these complex issues at the intersection of moral philosophy, critical theory, and gender studies to be able to provide context-specific assessments of algorithmic use. Aspects of these algorithms, all the way down to the very design decisions that produce them, are suffuse with normative implications. Thus, questions of their production, use, and evaluation belong properly under the purview of ethical theory.

We're then left with the question of how best to incorporate these varying perspectives in ways that allow experts in these intersecting disciplines to contribute to the development and use of technology. Drawing again on the comparison to scientific practice that has run throughout the paper, what this situation calls out for is the implementation of institutional mechanisms for ethical oversight for machine learning programs, perhaps through the extension of the Federal Drug Administration (FDA) or the creation of Institutional Review Boards (IRBs) for algorithms.[60] As Douglas (2014, 974) notes, "an IRB is another

[58] I express skepticism about statistical parity in the domain of race equality for two reasons. First, recall Castro (2019, 417-418)'s point that harms from the criminal justice system disproportionately affect members of the Black community. Second, I want to allow opportunities for preferential treatment toward historically disadvantaged groups so as to offset the accruement of historical injustices.

[59] COMPAS does not fare much better in this regard. As Angwin et al. (2016)'s analysis of COMPAS notes, "Surprisingly, given their lower levels of criminality overall, female defendants were 19.4 percent more likely to get a higher score than men, controlling for the same factors." For an extensive discussion about the relationship between gender and compounding wrongs, see Hellman MS.

[60] For an argument for the former, see **?**. I thank Tina Eliassi-Rad for pointing me to this suggestion.

mechanism for collectivizing responsibility" and "IRBs were put in place because scientists as individuals were making bad decisions, and societal anger was so strong, that IRBs were instituted to provide a check on the practices of scientists."[61] Regardless of the exact mechanism for the implementation of ethical oversight, what remains clear is the unacceptability of the status quo. That the explosion of technology has been allowed to grow unchecked and unfettered to the extent that it now permeates all facets of society without similar safeguards in effect is a profound injustice.[62] These are matters of research ethics that moral philosophers and policy-makers are well-positioned to contribute to.

Finally, to bring these normative reflections to a close, I want to draw attention to one final implication of the paper from the constitutive argument for demarcation. According to that argument, seemingly neutral features of an algorithm imbibe the social and political values that pervade the environment. The parasitic nature of algorithms on the world entail that certain morally objectionable qualities of the world will be picked up and perpetuated in the operation of the machine if left unchecked.

As mentioned at the start of the paper, people are well aware that algorithms can inherit these moral valences from the world through biases encoded in the data. The arguments of this paper extend these conclusions beyond the material data, to the material design of the algorithms themselves. Since our current social structures encode problematic and discriminatory patterns of oppression, programs that attempt to capitalize on those decision-making procedures that piggyback on the value-laden structure of the world will inherit problematic features. However, programs that abandon those patterns will fail to successfully make predictions about our present social environment. Thus, the adoption of machine learning programs for use in our present social environment necessarily involves an additional value-laden commitment about whether to inherit or abandon these problematic social patterns, prompting application of the justification argument against demarcation as well.

[61] Though, see discussion by Shiffrin (2014, 203-206) about the potential threats of IRBs to academic freedom.

[62] I would be remiss not to recognize here the tremendous work of grassroots initiatives that are well underway in attempting to provide such safeguards on the expansion and use of machine learning technology and big data, programs like the Algorithmic Justice League, the AI Now Institute, Data Feminism, Data for Black Lives, and FAT*ML. These initiatives make important contributions to the ultimate aim of keeping developers of social technology in check. However, they are naturally limited due to the lack of institutional authority that would guarantee compliance.

6 Conclusion

There's been a thematic undercurrent running throughout the discussions of this paper: inductive inferences generally, and machine learning programs specifically, are value-laden to the extent that they are connected to and dependent on matters that we care about as human beings. Unlike the detached and abstract formal logics of deductive inference, inductive inference depends constitutively for its risk, justification, and very nature on its connections to the world. The risks of getting algorithmic inferences wrong bring real-world consequences, their justification depends on patterns in the world continuing in familiar ways into the future, and the principles that govern their ability to traverse inductive gaps imbibe value structures manifest in society. This connection between value-ladennness as measured by impact on human endeavors has been noted before. As John Dupre (2007, 40) puts the point, "the fundamental distinction at work here is that between what matters to us and what doesn't." We might, in purely academic abstraction, be able to imagine a computer algorithm that operated wholly divorced from human endeavors. This algorithm, whatever it looks like, is far removed from the real-world technology that is the subject of this paper. So long as machine learning algorithms make important contributions to our daily lives, and the more we leave certain decision-making that affects the lives of individuals in the hands of algorithms, we must reject the value-free ideal of algorithmic decision making. Algorithms function to rank, sort, filter, categorize, assess, label, and draw any other number of conclusions about real-world phenomenon. They are the useful algorithms that they are only to the extent that they are undeniably value-laden.

7 Bibliography

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., and Robinson, D. G. (2020). Roles for Computing in Social Change. *Conference on Fairness, Accountability, and Transparency (FAT* '20)*, page 9.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Antony, L. (2001). Quine as Feminist: The Radical Import of Naturalized Epistemology. In Antony, L. and Witt, C. E., editors, *A Mind Of One's Own: Feminist Essays on Reason and Objectivity*, pages 110–153. Westview Press.

Antony, L. (2006). The Socialization of Epistemology. In Goodin, R. E. and Tilly, C., editors, *The Oxford Handbook of Contextual Political Analysis*, pages 58–77. Oxford University Press.

Antony, L. (2016). Bias: Friend or Foe? In Brownstein, M. and Saul, J., editors, *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, pages 157–190. Oxford University Press.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*.

Basu, R. (2018). The Wrongs of Racist Beliefs. *Philosophical Studies*.

Basu, R. (2019). What we epistemically owe to each other. *Philosophical Studies*, 176(4):915–931.

Bolinger, R. J. (2018). The rational impermissibility of accepting (some) racial generalizations. *Synthese*.

Bowker, G. C. and Starr, S. L. (2000). *Sorting Things Out: Classification and Its Consequences*. MIT Press.

Burge, T. (2020). Entitlement: The Basis for Empirical Warrant. In Graham, P. J. and Pedersen, editors, *Epistemic Entitlement*, pages 37–142. Oxford University Press.

Castro, C. (2019). What's Wrong with Machine Bias. *Ergo, an Open Access Journal of Philosophy*, 6(20191108).

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1610.07524*.

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, 4(2):205395171771885.

Corbett-Davies, S., Goel, S., and Gonzalez-Bailon, S. (2017). Even Imperfect Algorithms Can Improve the Criminal Justice System. *The New York TImes*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78.

Domingos, P. (2015). *The master algorithm: how the quest for the ultimate learning machine will remake our world*. Basic Books, a member of the Perseus Books Group, New York.

Dotan, R. (2020). Theory choice, non-epistemic values, and machine learning. *Synthese*.

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, 67(4):559–579.

Douglas, H. (2003). The Moral Responsibilities of Scientists (Tensions between Autonomy and Responsibility). *American Philosophical Quarterly*, pages 59–68.

Douglas, H. (2014). The Moral Terrain of Science. *Erkenntnis*, 79(S5):961–979.

Douglas, H. (2016). Values in Science. In Humphreys, P., editor, *Oxford Handbook in the Philosophy of Science*, pages 609–630. Oxford University Press.

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press, Pittsburgh, Pa. OCLC: ocn297144848.

Dupre, J. (2007). Fact and Value. In Kincaid, H., Dupre, J., and Wylie, A., editors, *Value-Free Science? Ideals and Illusions*, pages 27–41. Oxford University Press.

Fallis, D. and Lewis, P. J. (2016). The Brier Rule Is not a Good Measure of Epistemic Utility (and Other Useful Facts about Epistemic Betterness). *Australasian Journal of Philosophy*, 94(3):576–590.

Fantl, J. and Mcgrath, M. (2007). On Pragmatic Encroachment in Epistemology. *Philosophy and Phenomenological Research*, 75(3):558–589.

Flores, A. W., Bechtel, K., and Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation*, 80:38.

Friedman, B. (1995). Minimizing Bias in Computer Systems. *Mosaic of Creativity*, page 1.

Friedman, B. (1997). *Human Values and The Design of Computer Technology*. Cambridge University Press.

Friedman, B. and Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, 14:330–347.

Gardiner, G. (2018). Evidentialism and Moral Encroachment. In McCain, K., editor, *Believing in Accordance with the Evidence: New Essays on Evidentialism*, pages 169–195. Springer.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, 156(1):33–63.

Giraud-Carrier, C. and Provost, F. (2005). Toward a Justification of Meta-learning: Is the No Free Lunch Theorem a Show-stopper? *Proceedings of the ICML-2005 Workshop on Meta-learning*, page 8.

Gitelman, L., editor (2013). *"Raw Data" Is an Oxymoron*. MIT Press.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Bobbs-Merrill Indianapolis, Indianapolis.

Graham, P. J. and Pedersen, N. J. L. L., editors (2020). *Epistemic entitlement*. Oxford University Press, Oxford, first edition edition. OCLC: on1109918704.

Hellman, D. (2018). *Indirect Discrimination and the Duty to Avoid Compounding Injustice*. Hart Publishing.

Hellman, D. (2020a). Measuring Algorithmic Fairness. *Virginia Law Review*, 106:56.

Hellman, D. (2020b). Sex, Causation, and Algorithms: Equal Protection in the Age of Machine Learning. *Washington University Law Review*, 98:54.

Hu, L. and Kohler-Hausmann, I. (2020). What's Sex Got to Do With Fair Machine Learning? page 11.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Cambridge texts in the history of philosophy. Cambridge University Press, Cambridge ; New York. OCLC: ocm71808128.

Jeffrey, R. C. (1956). Valuation and Acceptance of Scientific Hypotheses. *Philosophy of Science*, 23(3):237–246.

Johndrow, J. E. and Lum, K. (2019). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220.

Johnson, D. G. and Nissenbaum, H., editors (1995). *Computers, Ethics, and Social Values*. Prentice-Hall, Inc.

Johnson, G. M. (2020a). Algorithmic bias: on the implicit biases of social technology. *Synthese*.

Johnson, G. M. (2020b). The Structure of Bias. *Mind*, 129(516):1193–1236.

Keas, M. N. (2018). Systematizing the theoretical virtues. *Synthese*, 195(6):2761–2793.

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

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180084.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, 165:74.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Kuhn, T. (1977). Objectivity, Value Judgement, and Theory Choice. In *The Essential Tension*. University of Chicago Press, Chicago.

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In Lakatos, I. and Musgrave, A., editors, *Criticisms and the Growth of Knowledge*, pages 91–195. Springer, Dordrecht.

Lauc, D. (2019). How Gruesome are the No-free-lunch Theorems for Machine Learning? *Croation Journal of Philosophy*, XVIII(54):8.

Levi, I. (1960). Must the Scientist Make Value Judgments? *The Journal of Philosophy*, 57(11):345.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Longino, H. E. (1995). Gender, politics, and the theoretical virtues. *Synthese*, 104(3):383–397.

Longino, H. E. (1996). Cognitive and Non-cognitive Values in Science: Rethinking the Dichotomy. In Hankinson Nelson, L. and Nelson, J., editors, *Feminism, science, and the philosophy of science*, pages 39–58. Kluwer, Dordrecht. OCLC: 801321444.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2020). A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635*, page 32.

Moss, S. (2018). *Probabilistic Knowledge*. Oxford University Press, Oxford.

Mulligan, D. K., Kroll, J. A., Kohli, N., and Wong, R. Y. (2019). This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–36. arXiv: 1909.11869.

Munton, J. (2017). The Eye's Mind: Perceptual Process and Epistemic Norms. *Philosophical Perspectives*, 31(1):317–347.

Munton, J. (2019a). Beyond accuracy: Epistemic flaws with statistical generalizations. *Philosophical Issues*, 29(1):228–240.

Munton, J. (2019b). Perceptual Skill And Social Structure. *Philosophy and Phenomenological Research*.

Nissenbaum, H. (1996). Accountability in a Computerized Society. *Science and Engineering Ethics*, 2:25–42.

Northpointe, I. (2015). *Practitioner's Guide to COMPAS Core*.

Norton, J. (2021). *The Material Theory of Induction*. BSPS Open.

Norton, J. D. (2003). A Material Theory of Induction. *Philosophy of Science*, 70(4):647–670.

Parker, W. S. and Winsberg, E. (2018). Values and evidence: how models make a difference. *European Journal for Philosophy of Science*, 8(1):125–142.

Ramsey, F. P. (1989). Mr Keynes on Probability. *The British Journal for the Philosophy of Science*, 40(2):219–222.

Rooney, P. (1992). On Values in Science: Is the Epistemic/Non-Epistemic Distinction Useful? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1992(1):13–22.

Rudner, R. (1953). The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science*, 20(1):1–6.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 59–68, Atlanta, GA, USA. ACM Press.

Shiffrin, S. V. (2014). *Speech matters: on lying, morality, and the law*. Carl G. Hempel lecture series. Princeton University Press, Princeton.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25(3):289–310.

Skeem, J., Scurich, N., and Monahan, J. (2020). Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and Human Behavior*.

Smith, A. (2018). Public Attitudes Toward Compuer Algorithms. *Pew Research Center*, page 41.

Stanley, J. (2005). *Knowledge and Practical Interest*. Oxford University Press, Oxford.

Steel, D. (2015). Acceptance, values, and probability. *Studies in History and Philosophy of Science Part A*, 53:81–88.

Tavernise, S. (2013). Drug Agency Recommends Lower Dosage of Sleep Aids for Women. *The New York Times*.

Titelbaum, M. G. (2010). Not Enough There There: Evidence, Reasons, and Language Independence. *Philosophical Perspectives*, 24(1):477–528.

Vidal, A. C., Smith, J. S., Valea, F., Bentley, R., Gradison, M., Yarnall, K. S. H., Ford, A., Overcash, F., Grant, K., Murphy, S. K., and Hoyo, C. (2014). HPV genotypes and cervical intraepithelial neoplasia in a multiethnic cohort in the southeastern USA. *Cancer Causes & Control*, 25(8):1055–1062.

Viens, L. J., Henley, S. J., and Watson, M. (2016). Human Papillomavirus—Associated Cancers—United States, 2008-20112. *MMWR Morb Mortal Wkly Rep*, (65):661–666.

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

Weisberg, M. (2007). Three Kinds of Idealization:. *Journal of Philosophy*, 104(12):639–659.

Wolpert, D. H. (2013). Ubiquity symposium: Evolutionary computation and the processes of life: What the no free lunch theorems really mean: How to imporve search algorithms. *Ubiquity*.

Zabell, S. (2011). Carnap and the Logic of Inductive Inference. In *Handbook of the History of Logic*, volume 10, pages 265–309. Elsevier.