Demarcating value demarcation in ML

Abstract

It has become widely recognized that machine learning (ML) systems are value-laden. This raises a value demarcation problem: how can we distinguish between legitimate and illegitimate non-epistemic value influences in ML development and use? This paper makes two contributions. First, it surveys value demarcation strategies in ML and identifies gaps in the debate. Second, it addresses a deeper issue: what makes for a good demarcation strategy? We need a way to judge the adequacy of existing demarcation strategies across contexts. I submit contextual adequacy as a meta-norm for evaluating the prima facie justification of value demarcation proposals in ML.

Introduction

In the context of machine learning (ML), the value-free ideal is dead and buried (Johnson 2023, Dotan & Milli 2019, Kraemer et al. 2011).[1] It is now clear that ML across the board is far from the politically neutral, epistemically objective tool that it was once thought it could be. Instead, throughout the whole ML research process, normative choices and tradeoffs are being made. Complete value-neutrality does not exist, and a crisp separation of epistemic and non-epistemic values also seems unattainable (Rooney 1992). Non-epistemic values come in through the back

[1] In philosophy of science, debates persist regarding the value-free ideal. However, within the domain of machine learning, there is widespread agreement that ML is inherently value-laden. I take this as a starting assumption and aim to make progress on what the implications are of accepting this claim.

door. Instead of leaving them to sneak around in the dark, we need to put a spotlight on this back door.

This is the value demarcation problem: if we accept that all systems are value-laden, how do we distinguish between legitimate and illegitimate non-epistemic value influences in ML development and implementation? This problem is especially pressing in social (i.e. non-scientific) applications, where the direct stakes of automated decisions are high for decision subjects. Individuals may have more or less well-grounded intuitions about acceptable and unacceptable biases in machine learning, but the solution to this hard problem should not be a matter of personal preference or intuition. A normative framework is needed to guide research standards and guidelines.

Previous work on the value demarcation problem falls within two broad categories: descriptive and normative. Descriptive work aims to explain the problem: its underlying reasons and what it looks like in practice, i.e. concrete value influences in the ML development process. There has been ample fundamental philosophical discussion about the value-ladenness of algorithms (Johnson 2023, Dotan & Milli 2019, Scheuerman 2021, Kraemer et al. 2011, Friedman 1996), fitting in a tradition of technology critique that goes back to Winner (1980), and there is also an increasing body of work on concrete value influences in practice (Birhane et al. 2022; Martin 2022; Ugar & Malele 2024; Biddle 2022; Angwin et al. 2016). Normative work, on the other hand, proposes a first-order criterion or set of criteria for value demarcation, such as providing reasons why a particular value is (il)legitimate.

This paper has two goals. First, drawing on discussions in philosophy of science, I categorize the current first-order norms that have been proposed and highlight open unexplored possibilities in the context of ML (section 2). Second, I identify the deeper problem of how we can evaluate why some demarcation norms are preferable to others, specifically in the context of ML. In other words, I introduce the need for meta-norms to be able to justify and evaluate first-order value demarcation strategies. I seek to make progress toward solving this deeper problem (section 3). Specifically, I suggest a norm for justifying what makes first-order norms (in)valid at first glance.[2] I argue that any strategy addressing non-epistemic values must result in ML that is at least contextually adequate (3.1). I then evaluate several current strategies along the meta-norm of contextual adequacy, with a close look on the value of fairness (3.2). Ultimately my aim is to push the debate about value-ladenness in ML forward from simply explaining that values are embedded in modeling, toward finding solutions to the hard problem of value demarcation.

2. Dominant first-order strategies: axiological and consequentialist

The AI/ML ethics landscape has been grappling with the value question for years. Taking stock, there are two dominant approaches regarding the influence of non-epistemic values: axiological strategies and consequentialist strategies.

2.1. Axiological strategies

Many call for appointing a certain value or set of values as (in)appropriate, often departing from philosophical ideas about e.g. social justice. These strategies draw the core distinction on the level of concrete values themselves. They establish clear norms for which particular non-epistemic values are allowed and which are not. Work that calls for concrete, single values

[2] Compare with ultima facie justification that asks finer-grained questions, such as how to solve conflicts between prima facie legitimate norms in practice.

such as fairness, transparency or accountability falls in this category (Hutchinson et al. 2021; Shah 2018; De Laat 2018; Zhou & Kantarcioglu 2020; Zerilli 2022; Nyrup 2022; Shook et al. 2017). Within these calls for concrete values, there are different views on what those values look like and how they should be achieved, resulting in different mitigation strategies. The details of implementation are not important here; what's important is to see that they all take the same axiological strategy.

> Axiological demarcation*:*
>
> Value *V* is (il)legitimate in ML.

Further examples of the axiological strategy include calls for decolonial (Mohamed et al. 2020; Mhlambi & Tiribelli 2023), feminist (D'Ignazio & Klein 2020; Toupin 2024; Hancox-Li & Kumar 2021), and anti-racist ML (Benjamin 2020; Buolamwini & Gebru 2018; Raji et al. 2020). These are not mutually exclusive and may be complimentary; indeed, calls for anti-racism often go hand in hand with calls for decolonial and feminist AI. They all seek to dismantle discriminatory values and promote values such as equity, inclusivity, anti-discrimination, and sometimes restorative justice.

2.2. Consequentialist strategies

An alternative first-order demarcation strategy takes as its starting point that as long as ML systems perform well, it does not matter which non-epistemic values shape our decision-making. The important thing here is the *consequences* values produce, rather than specific values themselves.

> Consequentialist demarcation:
>
> Consequence *C* determines which value(s) *V* are (il)legitimate in ML.

In ML, the most common perspective is that any value can shape decision-making as long as it does not impede model performance, often assessed through technical metrics; but other consequentialist views are possible. This performance-centric view is particularly prevalent among engineers (Birhane et al. 2022). However, as scholars have highlighted (Grote & Berens 2020; Danks & London 2017), we cannot escape the influence of and trade-offs between non-epistemic values in model development and use, from training data to methodological choices like model architecture, hyperparameters and objective functions. While standard loss functions aim to optimize for test-set accuracy, maximizing accuracy can sometimes conflict with other objectives such as fairness or privacy. Negotiating these trade-offs involves value-laden choices, even if they are not explicitly acknowledged. Thus, failing to explicitly consider non-epistemic value is a consequentialist stance, even if one may not be aware of it. Since we have moved past the value free ideal, not taking a stance is also taking a stance.

2.3. Gaps

These value demarcation strategies have their strengths and weaknesses. A major issue with axiological strategies for example is that we live in pluralistic societies with enormous variety in both non-epistemic values and ways of prioritizing them—and "it is unreasonable to assume that public deliberation will yield a shared set of values and a way of prioritizing them that is supported equally by all" (Holman & Wilholt 2022). Consequentialist strategies often fail to account for real-world harms or benefits (Birhane et al. 2022) and may open up (data) science to being abused for political gain (Holman & Wilholt 2022). Moreover, axiological and consequential strategies might conflict, e.g. when a certain value is deemed illegitimate yet results in great performance or vice versa.

Drawing on a current discussion in philosophy of science (Holman & Wilholt 2022), we can see that more strategies are possible. Axiological strategies locate the core distinction between legitimate and illegitimate value influences at the level of the values themselves, and consequentialist strategies locate it on the level of their effects. But there are other possible levels. First, functionalist strategies look at the role values take on in scientific research, not at what concrete values they are or the effect that they produce. Examples are Douglas (2009), who claims that for values to be legitimate they may act as reasons to determine evidence thresholds but are restricted from directly accepting or rejecting claims, preventing undue influence on the scientific process. Anderson (2004) argues that values are allowed as long as they do not "drive inquiry to a predetermined conclusion". Second, coordinative strategies legitimize value-laden methodological choices by their alignment with expectations placed on them by others. For example, those choices should adhere to certain standards (and meta-standards) in order to be legitimate (Elliott & McKaughan 2014). Finally, systemic strategies make the demarcation dependent on the whole social set-up within which the research occurs: the whole system needs to meet certain conditions for value-influences in science to be legitimate (Longino 1990, Kitcher 2011). An example of a systemic strategy for ML is Paullada et al. (2021), who argue that a turn in the culture is necessary regarding dataset development and use. Note that these strategies are not strictly mutually exclusive and that some may often produce coinciding demarcations in practice. For example, a decolonial strategy is mainly axiological since it locates the core distinction between legitimate and illegitimate value influences on the level of values themselves, but it has elements of a systemic strategy; working towards decolonial ML is not a matter of picking out bad apples, but changing the entire system.

There are two central gaps with corresponding avenues for further research. One gap is that not all strategies have been explored in detail in the context of machine learning. One possible route to take is thus to establish more sophisticated first-order demarcation strategies by drawing parallels with similar views in philosophy of science. For example, one could take a functionalist or coordinative approach and apply it to the machine learning context. This is certainly fruitful and would advance our understanding of non-epistemic values in machine learning. The second gap however is more interesting and more challenging. There is a lack of work that provides normative guidance on which of those strategies are useful, empirically successful, or normatively grounded and justified. What makes for a good demarcation strategy? Why should some non-epistemic value influences prevail over others? While some obvious candidates come to mind: goals might be to minimize discrimination or human suffering, to improve well-being or equity, these goals may conflict with other goals, e.g. seeking truth or driving innovation. Thus, in order to avoid justification halting at "just-so" and ending up with relatively static, possibly inconsistent, or generally unclear first-order norms, higher-level goals of ML require careful consideration. In this paper, I seek to make progress on this second research gap by proposing contextual adequacy as a meta-norm for prima facie justification of value demarcation strategies in ML.

3. Contextual adequacy as a meta-norm

Establishing higher-level goals of ML is no easy feat. Ultimately the question boils down to: what do we want machine learning to be? In philosophy of science, there are hotly debated goals and ideals that science as a practice should live up to. Understanding the natural world, prediction and control, generalizability, accuracy, and impartiality or variations thereof come to mind. But there are also voices in the debate that argue it may not be possible to come up with overarching umbrella goals for science, and instead we need to look at concrete disciplines. I am sympathetic to the latter view. My goal here is not to characterize ML in its entirety and arrive at an exhaustive list of meta-norms that can provide the solution to all issues with first-order proposals in practice. It is much more modest: the aim is to propose contextual adequacy as a meta-norm for ML, which can be used to make an initial evaluation of demarcation strategies in non-scientific contexts.

3.1 Adequacy for purpose (AFP)

ML models are usually developed for specific tasks in specific contexts. However, they are often adapted and used in contexts beyond the context of development. For example, a ML system that was trained to classify pastries learned to perform a variety of other tasks, including the identification of cancer cells (Somers 2021). Transferability is particularly valuable when confronted with challenges such as limited training data or resource constraints: learned features can be generalized to new models or contexts, thereby reducing the necessity to start training models from scratch for each unique application (Ching et al. 2018). Moreover, transferable models are more robust against deviations from assumptions made in model training, i.e. models will be better equipped to handle variations in real-world scenarios (Zhuang et al. 2020, Weiss et al. 2016).

There are technical trade-offs in the quest for transferability. One highly effective approach is fine-tuning: models inherit pre-trained weights from another model and start learning from there, as opposed to random initialization, where the initial values of the model's parameters are set randomly. Fine-tuning greatly improves accuracy (Yosinski et al. 2014). However, it has been found that vulnerability to adversarial attacks transfers too; attacking the pre-trained model will deceive the transferred model (Rezaei & Xin 2019). Models trained with random initialization are much more robust against such attacks, though they exhibit lower accuracy (Chin et al. 2021). Different transfer learning methods are proposed to deal with this problem. This is a fruitful avenue in computer science.

Beyond technical trade-offs, however, transferability comes with certain risks and harms. ML is not only a technical endeavor; it is so embedded in social structures that these provide constraints on transferability too. It is crucial to judge whether a model is actually desirable in a new context of application, even if it is technically transferable. In other words, judging transferability is not just a matter of accuracy and robustness against dataset shifts. For example, implementing a medical decision support system in a new context that has demonstrated high accuracy across datasets and robustness against adversarial attacks may still result in adverse consequences; doctors in certain regions may put either too little or too much trust in systems, or there may be a cultural emphasis on the patient-doctor relationship, and people may feel scared of or misunderstood by a "machine". In general, research has suggested that ML in healthcare disrupts existing work practices, which disturbs patient-doctor or patient-nurse relationships (Elish & Watkins 2020). Different application contexts have different requirements, partly determined by non-epistemic values, that go beyond performance metrics. Thus a consensus has emerged that ML should be contextually adequate; not taking the context into account can result in various harms.

However, the details of such contextual adequacy have not been fleshed out. What makes a model adequate for purpose and how can we assess this in practice? Here I draw inspiration from Parker (2020), who developed an adequacy-for-purpose (AFP) view of evaluation for scientific models. On the AFP view, models cannot be evaluated as good or bad just by representational accuracy. Instead, models should always be assessed in light of their purpose or downstream use. The quality of the model therefore only exists relative to a certain application context.

In order to judge whether a model is adequate-for-purpose, one first needs to understand the purpose and how that purpose can be met. Note that in non-scientific contexts, the purpose of a ML model is not always epistemic. Models can be used for a range of goals, such as resource allocation, customer service, cost minimization, automation of several tasks, etc. It may often be the case that the purpose is achieved through epistemic sub-purposes, but not always. And whether a purpose is achieved is often open to interpretation; if the goal is to explain a certain event, it depends on your notion of explanation whether that goal has been met successfully. In many cases the purpose needs to be defined more clearly in the first place, and sometimes there are reasons to reject the purpose outright. For example, in the case of people using Meta's open-source LLM to create graphic sexbots that engage in violent and illegal acts, we have good reasons to argue that no model would be adequate for this particular purpose (Dupré 2023).

If the purpose of the model is clear and acceptable, we need ways to evaluate the adequacy of a certain ML model for that purpose, i.e. whether people can actually achieve that purpose through the use of that model. The key insight is that for a ML model to be AFP, it needs to be not only accurate and robust against adversarial attacks (i.e. stand in the right relationship with the target system) (T), but it also needs to stand in the right relationship with the user of the model (U), the methodology that is employed (W), and background assumptions that are at play (B) all at the same time. In Parker's words, purpose P defines a problem space and U, W, B, T are constraints on how it should be solved. The model M then needs to be a solution within that problem space. These constraints interact; sometimes T *depends on* U, W, or B. Take a complex phenomenon A that you want to explain with model M; the purpose of the model is thus to explain A. If the target is represented in too much detail, it may result in a model that is incomprehensible to users (U) and therefore fails to explain P. In this case, what the suitable relationship to the target is depends on user needs and constraints. Note that the constraints are not just epistemic; there are non-epistemic value constraints at play in U, W, B and T (Lusk & Elliott 2022).

Thus, the upshot of the meta-norm of contextual adequacy is that the values that are encoded in ML systems should be brought to light and scrutinized whether it is a good fit for a particular purpose. Systems cannot be rejected or endorsed in the abstract, evaluation should always happen in light of a particular application. To this end, extensive documentation is crucial, which can be standardized by e.g. model cards (Mitchell 2019). An account of legitimate value influences in ML should specify how it produces systems that are adequate for purpose. Practical questions include: What is the purpose of this model? When can we say the purpose is achieved? Are the application constraints of fidelity, user characteristics, methodology, and background context specified? Is there a clear roadmap for when models do or do not meet the application constraints?

3.2 Evaluation

Let us evaluate the dominant first-order strategies seen in section 2. Consequentialist strategies maintain that any system is transferable to any context as long as it performs well on certain metrics. This clearly does not produce ML that is contextually adequate; it does not specify in any way how application constraints of user characteristics, methodology, and background context should be defined, and how models can meet them.

We have already seen an example with the medical decision support system, where the model was performing well but was not adequate for purpose in a different application context due to cultural user constraints. Examples of where M is not a solution in the problem space constrained by U, W, B, T but are still used are plenty. ChatGPT is used for customer service interactions in diverse multinationals, which is not adequate for purpose due to non-epistemic value influences related to cultural sensitivities and nuanced language understanding (Shrivastava 2023). Or consider a sentiment analysis tool for academic recruitment (Uloko et al. 2023). The idea is that sentiments in people's resumés can be analyzed and leveraged for better hiring decisions. Text like "I am a highly motivated individual…" is labeled positive, whereas text like "I am interested in a software development role…" is labeled neutral. Each applicant then gets an aggregated sentiment score, which, according to the researchers, enhances "the institution's ability to make well-informed decisions that encompass both eligibility and suitability aspects" (Uloko et al. 2023). From a consequentialist perspective, the primary focus might be on the tool's ability to accurately predict sentiment and its impact on hiring success rates. However, such a tool might inadvertently introduce biases, potentially disadvantaging candidates from certain demographics or cultural backgrounds.

The meta-norm of contextual adequacy allows us to see that we can outright reject the dominant consequentialist views that implicitly argue that any value influence is allowed as long as it does not hinder performance.[3] Contextual adequacy involves a holistic understanding of societal, cultural, and ethical dimensions involved in a certain purpose, beyond narrow performance metrics. This is a less trivial point than it seems, given the still widespread attitude among computer scientists that performance is the primary objective and anything resembling ethics is an afterthought. It also forces us to consider the purpose of systems more often, and whether we should use the system in the first place.

[3] Note that this does not imply that *all* consequentialist views suffer the same fate.

Some axiological strategies seem to fare better here; others do not. Not all value demarcation proposals align with the meta-norm of contextual adequacy. Let us take fairness as an example. There is a narrow understanding of fairness as a statistical notion that is obviously unhelpful when aiming to achieve contextual adequacy for ML. When trying to express fairness in technical metrics, it is clear that there are many different options, and some definitions of fairness are undesirable in certain contexts. Demographic parity for example aims to ensure equal outcomes across different demographic groups, which may not be desirable in medical diagnosis and treatment. Certain diseases may affect demographic groups differently, and treatment should be based on clinical factors rather than demographics alone. Thus, achieving fairness in healthcare is not simply a matter of achieving demographic parity. A value demarcation proposal that stops at simple definitions of statistical fairness therefore will not be useful in our quest for (de)legitimizing non-epistemic value influences in ML, since such definitions do not facilitate contextual adequacy.

It has been frequently pointed out that a static axiological approach results in the impossibility of fairness; there are many statistical fairness definitions that are impossible to all optimize at the same time, and sometimes there are multiple desirable definitions of fairness that conflict (Chouldechova 2017; Friedler et al. 2021). This seems to put us in a bind: satisfying one definition of fairness will violate at least one other desirable definition of fairness. Algorithmic fairness, in other words, does not always lead to justice in practice (Hoffmann 2019; Binns 2018; Selbst et al. 2019). There are different responses to this problem. Some argue that, in fact, there is a metric that trumps others (Dwork et al. 2012, Corbett-Davies et al. 2017); some argue that the impossibility problem is a theoretical problem but not a practical one (Bell et al. 2023);

others have called for a rejection of the (axiological) fairness frame altogether, and towards other approaches for value demarcation.

Green (2022) for example proposes methodological reform. This is a combination of a systemic and axiological strategy for value demarcation, since it maintains that fairness is legitimate only if certain conditions in the system have been met. He argues that the higher-level goal of many ML applications is to promote justice. The focus on formalization of fairness results in the impossibility problem; instead, the development and use of algorithms should draw on philosophical theories of substantive equity. Instead of treating fairness as a technical attribute of algorithms, Green maintains, we need to look at how algorithms promote justice in practice. His approach involves two steps: reducing upstream social disparities that feed into decision-making processes, and reducing downstream harms for those disadvantaged within decision-making processes (Green 2022: 4). These complementary steps appeal to the conditions of the sociotechnical system as a whole, not just to values themselves or the effects thereof.

Using the meta-norm of contextual adequacy, we can see that this approach is preferable over others. The purpose is clear: promote social justice in practice. For this purpose to be met, then, certain conditions in the broader socio-technical system need to be eliminated (B). This approach to algorithmic fairness is supposed to account for relational and structural harms by eliminating certain conditions in the broader socio-technical system (B) through methodological reform (W), such that users (U) of the model (also including decision subjects) can achieve the purpose (P) of promoting social justice in practice. The proposed methodology is one of “substantive algorithmic fairness”, which adopts tools from the discussion on substantive equality “to reason about when formal algorithmic fairness is (and is not) appropriate” (Green 2022: 16). Though there is much more to be said here, it is clear that this systemic and axiological approach is better equipped to achieve contextual adequacy than purely axiological approaches to value demarcation. The upshot is that any approach that calls for value demarcation without specifying the purpose of a system and how we can determine whether the system is adequate for that purpose can be rejected outright, or should at least be accommodated.

Unfortunately there is no space here to consider other value demarcation proposals. However, I hope that this gives enough inspiration to see how the meta-norm of contextual adequacy might prompt deeper consideration of the purposes of ML applications and provide guidance on how to evaluate value demarcation proposals. It can highlight what value demarcation proposals should take into account, explain how they can be improved, or maybe even provide structured reasons for why they can be rejected outright.

Conclusion

Non-epistemic value influences are pervasive throughout the ML lifecycle. This problem has invited many accounts of (il)legitimate values in ML, most notably axiological and consequential strategies. These accounts might conflict and other strategies (functionalist, coordinative, systemic) are possible, which brings us to the question of justifying certain strategies over others.

I have argued that in order to make progress on this value demarcation problem, meta-norms need to be established based on the higher-level goals of ML research. Our suggestion is that one such higher-level goal of ML is contextual adequacy. This meta-norm can be used to evaluate first-order value demarcation proposals.

Note that I have taken only a small step. Our aim was not to provide an exhaustive list of meta-norms, nor a comprehensive framework for what we should do all things considered. Other meta-norms will have to be thought out and conflicts may occur. However, if we hope to ever make progress on the issue of value-laden machine learning, I contend that at least we can now see that this is the right direction.

References

Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2022). Contestable AI by Design: Towards a Framework. *Minds and Machines*, 1-27.

Almada, M. (2019, June). Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (pp. 2-11).

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia*, *19*(1), 1-24.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias risk assessments in criminal sentencing. *ProPublica, May*, *23*.

Benjamin, R. (2023). Race after technology. In *Social Theory Re-Wired* (pp. 405-415). Routledge.

Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., & Stoyanovich, J. (2023, June). The possibility of fairness: Revisiting the impossibility theorem in practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 400-422).

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

Binns, R. (2018, January). Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency* (pp. 149-159). PMLR.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022, June). The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 173-184).

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Chin, T. W., Zhang, C., & Marculescu, D. (2021). Renofeation: A simple transfer learning method for improved adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3243-3252).

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., ... & Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the royal society interface*, *15*(141), 20170387.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, *20*(4), 336-349.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. Proceedings of the 23rd ACM SIGKDD International Conference on Knowl- edge Discovery and Data Mining. https://doi.org/10.1145/3097983.3098095

Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. In *Ijcai* (Vol. 17, No. 2017, pp. 4691-4697).

De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?. *Philosophy & technology*, *31*(4), 525-541.

D'ignazio, C., & Klein, L. F. (2023). *Data feminism*. MIT press.

Dotan, R., & Milli, S. (2019). Value-laden disciplinary shifts in machine learning. *arXiv preprint arXiv:1912.01172*.

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

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of science*, *67*(4), 559-579.

Dupré, M.H. (2023). People Are Using Meta's New AI to Make Graphic Sexbots. In: Futurism. https://futurism.com/metas-new-ai-graphic-sexbots

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic values and the multiple goals of science. *Philosophy of Science*, *81*(1), 1-21.

Elish, M. C., & Watkins, E. A. (2020). Repairing Innovation: A Study of Integrating AI in Clinical Care (p. 62). Data & Society. https://datasociety.net/library/repairing-innovation/

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on information systems (TOIS)*, *14*(3), 330-347.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, *64*(4), 136-143.

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, *46*(3), 205-211.

Hancox-Li, L., & Kumar, I. E. (2021, March). Epistemic values in feature importance methods: Lessons from feminist epistemology. In *proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 817-826).

Henin, C., & Le Métayer, D. (2021). Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY*, 1-14.

Hirsch, T., Merced, K., Narayanan, S., Imel, Z. E., & Atkins, D. C. (2017, June). Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (pp. 95-99).

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900-915.

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, *2*(4).

Holman, B., & Wilholt, T. (2022). The new demarcation problem. *Studies in history and philosophy of science*, *91*, 211-220.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Johnson, G. M. (2023). Are Algorithms Value-Free?: Feminist Theoretical Virtues in Machine Learning. *Journal of Moral Philosophy*, *1*(aop), 1-35.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton university press.

Lusk, G., & Elliott, K. C. (2022). Non-epistemic values and scientific assessment: an adequacy-for-purpose view. *European Journal for Philosophy of Science*, *12*(2), 35.

Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-25.

Kotliar, D. M. (2020). Data orientalism: on the algorithmic construction of the non-Western other. *Theory and Society*, *49*(5-6), 919-939.

Kourany, J. A. (2022). The new worries about science. *Canadian Journal of Philosophy*, *52*(3), 227-245.

Kraemer, F., Van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms?. *Ethics and information technology*, *13*, 251-260.

Intemann, K. (2015). Distinguishing between legitimate and illegitimate values in climate modeling. *European Journal for Philosophy of Science*, *5*, 217-232.

Martin, K. (2022). Algorithmic bias and corporate responsibility: How companies hide behind the false veil of the technological imperative. In *Ethics of data and analytics* (pp. 36-50). Auerbach Publications.

Mhlambi, S., & Tiribelli, S. (2023). Decolonizing AI ethics: Relational autonomy as a means to counter AI Harms. *Topoi*, *42*(3), 867-880.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, *33*, 659-684.

Nyrup, R. (2022). The Limits of Value Transparency in Machine Learning. *Philosophy of Science*, *89*(5), 1054-1064.

Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, *87*(3), 457-477.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11).

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020, February). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 145-151).

Rezaei, S., & Liu, X. (2019). A target-agnostic attack on deep models: Exploiting security vulnerabilities of transfer learning. *arXiv preprint arXiv:1904.04334*.

Rooney, P. (1992). On values in science: Is the epistemic/non-epistemic distinction useful?. In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1992, No. 1, pp. 13-22). Cambridge University Press.

Scheuerman, M. K., Hanna, A., & Denton, E. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-37.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59-68).

Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170362.

Shook, J., Smith, R., & Antonio, A. (2017). Transparency and fairness in machine learning applications. *Tex. A&M J. Prop. L.*, *4*, 443.

Shrivastava, R. (2023). ChatGPT is coming to a customer service chatbot near you. https://www.forbes.com/sites/rashishrivastava/2023/01/09/chatgpt-iscoming-to-a-customer-service-chatbot-near-you/

Smart, A., James, L., Hutchinson, B., Wu, S., & Vallor, S. (2020, February). Why reliabilism is not enough: Epistemic and moral justification in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 372-377).

Somers, J. (2021). The Pastry AI That Learned to Fight Cancer. *The New Yorker*, *18*.

Toupin, S. (2024). Shaping feminist artificial intelligence. *New Media & Society*, *26*(1), 580-595.

Ugar, E. T., & Malele, N. (2024). Designing AI for mental health diagnosis: challenges from sub-Saharan African value-laden judgements on mental health disorders. *Journal of Medical Ethics*.

Uloko, F., Enihe, R. O., & Obrorindo, C. I. (2023). A Sentiment Analysis Based Model for Recruitment by Higher Institutions. *Journal of Computer and Communications*, *11*(9), 44-56.

Vaccaro, K., Karahalios, K., Mulligan, D. K., Kluttz, D., & Hirsch, T. (2019, November). Contestability in algorithmic systems. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 523-527).

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*, 1-40.

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science Part A*, *40*(1), 92-101.

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136.

Yang, Z. (2019, July). Fidelity: A property of deep neural networks to measure the trustworthiness of prediction results. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security* (pp. 676-678).

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. *Advances in neural information processing systems*, *27*.

Zhou, Y., & Kantarcioglu, M. (2020). On transparency of machine learning models: A position paper. In *AI for Social Good Workshop*.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43-76.