‘The purpose of a system is what it does, and science is a thing which people do’: from AI epistemology to AI military ethics

Zhanpei Fang
Departments of Geography & EECS
Oregon State University
Corvallis, OR 97331
fangzha@oregonstate.edu

1 Introduction

Drawing upon recent philosophical work on understanding the usage of machine learning for natural sciences research, I posit that the epistemic problems of machine learning have significant bearing upon concerns related to its deployment in a military context. In particular I try to sketch out the throughlines between the epistemology and the ethics of AI by way of the useful philosophical lenses of the *theory-free ideal* and *instrumental reason*. The urgency of this task is underlined when we consider ML/AI’s growing role in administering human life, in military and statecraft as well as in many other contexts.

The faulty epistemic practices performed by AI practitioners, commentators and policymakers have real consequences on the social and natural world. I consider the ethical consequences of the ‘conceptual poorness’ assigned to machine-learning methods, and provide some theoretical scaffolding which might allow AI to be folded into other discourses of technology, namely the critique of instrumental reason, additionally applying the Marxian notion of reification to the task of understanding AI as a *social technology* or organizing activity. Informed by my experiences as a junior applied-ML researcher in the space-tech industry, and now in academia studying novel deep-learning methods on satellite imagery in regions of humanitarian & conflict concern, and illustrating with some recent examples[1], I provide a few policy recommendations as well as recommendations for AI ethics & fairness research directions.

2 Summary of argument

2.1 The theory-free ideal, in basic sciences research and beyond

Mel Andrews, addressing its uptake in the context of ML’s widespread adoption in the natural sciences, explicates the ‘theory-free ideal’ as the meta-narrative that ML exists as a theory-free enterprise, on a novel and disruptive epistemic footing relative to classic statistical approaches, leading to a belief that “science will undergo drastic change with the advent of ML-based methods, because such methods are theoretically unmoored or conceptually impoverished in a way that sets them fundamentally apart from existing methods” [1]. This is (1) a *disruption* claim, that ML will radically disrupt scientific practice, which is founded on (2) a *distinctness* claim. Taking apart this ‘theory-free ideal’, which they claim has a “deleterious effect on the epistemic standing of ML-based science", Andrews argues that ML models and the data gathered for them are ‘theory-laden’ by virtue of their provenance,

[1]Not included for the purposes of the extended abstract, but will be included in poster to discuss recent usages of AI for military-target identification.

2024 Harms and Risks of AI in the Military (HRAIM 2024).

processing and interpretation, and that 'raw' data never 'speaks for itself'. I emphasize that the same applies for ML deployed in other contexts and *their* downstream goals.

Hogg & Villar, who identify themselves as natural scientists, recently presented an ICML 2024 position paper on "Is machine learning good or bad for the natural sciences?" [2] wherein they posit that ML methods have a *strong ontology* (in which only the data exist) and *strong epistemology* (in which a model is considered good if it performs well on held-out training data)–which they claim is in strong conflict with standard practices and key philosophies in the natural sciences, wherein we care about understanding the world. This issue is related to the philosophical differences which have been posited to exist between frequentist and Bayesian statistics (see [3] for a practitioner's perspective), the latter of which's epistemology is typically concerned with 'degrees of belief' and credence changes; see [4], particularly §1.9 and §6 on idealization as it applies to scientists' use of Bayesian statistics for science.

The technical [5] and philosophical challenges of uncertainty quantification in the machine-learning context, which heavily relies upon Bayesian inference, are value-laden. Andrews warns against the "'laundering' of uninterrogated values into the outputs of such ML-based decision-making and decision-support systems, where they are then reified as objective empirical truth"; I argue that value-decisions made when applying ML in the social domain are especially worthy of extreme caution.

2.2 Against the (Marxist) reification of artificial intelligence

'Reification' is a useful term with which to capture the 'magical' or 'existential' color which people often give to AI. Andrews' above usage of 'reification' differs only in shades from the Marxist definition of the word, proposed by Lukács [6] and formalized by Adorno in *Negative Dialectics* [7] (in whose interpretation I am gratefully assisted by Gillian Rose's *The Melancholy Science* [8]). The latter school characterizes reification as a set of social relations in the false appearance of concrete form. As I elaborate upon in §2.3-3, technology may be understood as something which is *reified*, "appear[ing] as independent beings endowed with life, and entering into relation both with one another and the human race" [9].

AI research is built upon very human models of how we conceive of and do science, which then present a diverse collection of reified concepts. I argue that ML modeling used in a military capacity which engages 'social' data, for example a supervised learning model given human-annotated training data for target-identification decision support, in particular concretizes social relationships into forms which can be quantitatively measured and optimized. The AlphaFold example Andrews describes sits upon a wealth of biology *domain knowledge*–which likewise applies to the 'human' systems AI has been used to describe and optimize as well, but in the latter case we might more appropriately call it a reified image of society, or even perhaps *false consciousness*.

2.3 Machine learning as instrumental reason

Here I try to characterize ML's ideology as a form of instrumental reason by way of critical-theoretic and philosophical accounts of technology, in particular drawing upon the thought of Max Horkheimer and Martin Heidegger. Horkheimer's theory [10, 11] of the *instrumental mode of reason*, which existed in pre-capitalist society but only became a 'structuring principle' in capitalism, and from which positivist science flows, is particularly well-suited to describing AI. He argues that in modernity the concept of reason has been reduced to an *instrument* for achieving practical goals assessed on its operational value, rather than a means of understanding objective truth–a kind of thinking which emphasizes ends rather than the means employed. He claims that the inexorable drive of instrumental reason results in a distorted picture, which is falsely understood as the only true picture of the world[2].

Horkheimer's framework is concordant with the one proposed by Heidegger in "The Question Concerning Technology" [12], wherein he claims that "the essence of technology is by no means anything technological", and similarly identifies technology as a means to a human end, coupling its *instrumental* and *anthropological* definitions: "For to posit ends and procure and utilize the means to them is a human activity." This is misleading, however, because it encourages us to think that "by making the technology better–better able to 'get things done'–we will master technology

[2]As Adorno would have characterized it, a form of *identity thinking* [7, 8].

and solve the problems that accompany it." In the latter half of the essay Heidegger goes into a discussion of 'cause' and formulation of technology as a kind of *poesis*, a way of bringing forth or revealing; but argues that modern technology's mode of revealing is not *poesis* but a 'challenging forth' which transforms our orientation to the world (enframing, *Gestell*), converting the natural world and humanity itself to some extent to 'standing reserve'–an argument which can be mapped to Horkheimer's. Technology involves dominating outer nature for human purposes, but also making over society for human purposes.

The mode of reason encouraged by the manner in which artificial intelligence has historically developed as a research and an applied discipline may in itself be considered as a Heideggerian enframing, an orientation towards the world. Hogg & Villar's perspective[3] as natural scientists from earlier can perhaps be echoed by a (strong, perhaps vulgar) claim that ML is a crystallization of the instrumental reason which characterizes modernity [11]. The dimension of means-end in instrumental reason echoes the ideological generator behind machine learning, wherein the final model *prediction* is prioritized above the model architectures we used to get there, optimizing an objective function by whatever means necessary–which may very well also be the guiding ideology of research into convex [13] and non-convex optimization, at least in some of its applied forms.

3 Conclusions and recommendations

The idea of an ML model being an inexplicable black-box because "there are no *a priori* assumptions concerning the mechanism of the target phenomenon" [1] is very dangerous in the context of its military usage. I argue that AI is not necessarily novel *sui generis*; this lack of novelty allows us to apply other discourses and philosophies of technology from previous iterations of techno-warfare and bureaucratized pretexts for violence. The competency or fidelity of these systems is besides the point, and increased human oversight or explainability in themselves may not solve the problems of militarized AI which we would like them to. What is somewhat novel is the perceived opaqueness of these models, which creates a fog of epistemic indeterminacy mimicking the 'fog of war', in which atrocities may perpetuate; and creates room for the slippage of intention, even when those intentions are explicitly baked in from the start.

AI can be understood as a *social technology*, by which I mean the social arrangements and management methods which are themselves technologies that propagate and alter social life, just as much as the models they muster data for. This social technology hails [14] vast material resources, finance capital, and institutional infrastructure, with its own supply chains and political economy. What we should be concerned about are the social arrangements which allow human or natural data to be interpellated or called forth to serve such ends. In thinking about militarized AI we must consider the preconditions of surveillance required to conceptualize and gather data for these models, and their stated technological or security justifications. Our task is to work out a critique which goes beyond simply calling for more explainability research or for more 'human-in-the-loop'; because humans are already incredibly present in the loop for any model, from its conceptualization to data collection, preparation, manipulation, and interpretation.

The extent to which these 'military AI systems' are credible or actually used may in some ways be irrelevant, because the main purposes they serve are ideological or rhetorical, with massive psychological benefits for those pressing the buttons, and shifts the focus from the social relations between people to the technologies used to implement them, a mystification which misdirects focus and propagates invincibility. By being preoccupied by the purported technical intricacies and obscurities of automated war-making methods one fetishizes them, precluding substantive critique or action.

In this note I have begun exploring the throughlines between AI epistemology and AI ethics; and how faulty epistemic assumptions may then lead to faulty ethics. One priority is to not allow the space of uncertainty opened by AI's epistemic problems to go unclosed and unaddressed. A call to action might be to think about these epistemic problems explicitly in the context of deciding 'best practices' and disarmament–in policy as well as in evaluating funding proposals—treating these projects with far more credulity with respect to their epistemic soundness. This might take the form of moderate adjustments to or even perhaps deep structural changes in publication or funding incentives.

[3] "A ML method is considered successful if it performs well on held-out training data, even if the latent structure of the model is generic and the internals are impossible to predict" [2].

I further propose that we try to understand the term 'artificial intelligence' in some of the flavors of its usage as a social-rhetorical formation which produces justifications for conclusions or goals already determined in advance; highlighting its capacities as an intangible social technology and faulty rhetorical mode which elides human intention, creating the space of epistemic indeterminacy through which people act.

References

[1] Mel Andrews. The Immortal Science of ML: Machine Learning & the Theory-Free Ideal. 06 2023. URL http://dx.doi.org/10.13140/RG.2.2.28311.75685/1.

[2] David W. Hogg and Soledad Villar. Is machine learning good or bad for the natural sciences? In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2405.18095.

[3] Jake VanderPlas. Frequentism and Bayesianism: a Python-driven primer. *Proc. of the 13th Python in Science Conf. (SciPy 2014)*, 2014. URL https://arxiv.org/abs/1411.5018.

[4] Hanti Lin. Bayesian Epistemology. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition, 2024.

[5] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.05.008. URL https://www.sciencedirect.com/science/article/pii/S1566253521001081.

[6] Georg Lukács. *History and Class Consciousness: Studies in Marxist Dialectics*. MIT Press, 1972.

[7] Theodor Adorno. *Negative Dialektik*. Suhrkamp, 1966.

[8] Gillian Rose. *The Melancholy Science*. Springer, 1978.

[9] Karl Marx and Friedrich Engels. *Capital*. Number 1 in Capital. Progress Publishers, 1887. URL https://www.marxists.org/archive/marx/works/1867-c1/.

[10] Max Horkheimer. *Eclipse of Reason*. Continuum, New York, 1947.

[11] Max Horkheimer and Theodor W. Adorno. *Dialectic of Enlightenment: Philosophical Fragments*. Stanford University Press, Stanford, Calif., 2002.

[12] Martin Heidegger. *The Question Concerning Technology, and Other Essays*. Harper & Row, New York, 1977.

[13] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[14] Louis Althusser. *Lenin and Philosophy, and Other Essays*. Monthly Review Press, New York, 1971.