

Workshop: Values in science and machine learning (list of abstracts)

Kathleen Creel: Does machine learning model convergence support value alignment?

The Platonic Representation Hypothesis posits that “neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces” (Huh et al. 2024). This realist explanation for the success of contemporary large language and image models has gained traction within computer science, particularly as an explanation for some of the striking large model convergence results found using mechanistic interpretability techniques. In this talk, I undermine the Platonic Representation Hypothesis both by positing alternative explanations for the empirical phenomenon of model convergence and by undermining the value-free ideal at which model convergence aims.

Nico Formanek: Is every decision in ML ethical?

A popular argument why ML methods must encode ethical decisions relies on two key assumptions. First, there must be some unwanted moral effect of the ML method, and second, there must be some way of attenuating this effect. One then works backward from the effect to the decisions that caused it. If these decisions are then unconstrained by data, implying freedom of choice, then whoever made them could have done differently. One problem with such arguments is that, apart from simple inductive risk style cases, the consequences of many hyperparameter settings are not clear. Furthermore, even if negative consequences do appear, they probably were unintentional, posing not only a problem for the attribution of blame, but also for what should have been done instead. In short, we have a problem of separating ethical from “mere” technical decisions. One radical answer to this problem is denying the distinction. Such denial is not without precedent, Dinger, for example, held that science is founded in ethics, that it is fundamentally ethical. In my talk I will explore this seemingly radical solution for the case of ML methods. One ramification is immediately visible, every decision in ML is ethical.

Carina Hausladen: Alignment for Whom? Choices about Bias and Welfare in AI

What does it mean to build AI that is fair, aligned with human values, and supportive of democracy?

I approach this question in three steps: how we measure bias, whose preferences we prioritize, and how we translate democratic values into computational systems.

First, I examine how bias in AI is measured, drawing on the social sciences’ long tradition of studying discrimination, from audit studies to emerging computational approaches. Measuring bias is itself a choice: what we measure, where we measure it in a model, and against which standard all shape what we find. I present a causal approach to measuring bias in vision-language models.

Second, I turn from measuring bias to shaping AI around human values. Learning from human feedback offers one way to incorporate human preferences, but whose preferences should guide AI when people disagree? Drawing on research on alignment preferences and social choice, I show how different ways of aggregating human preferences lead to different conceptions of whose welfare AI should serve.

Finally, I examine the opportunities and challenges of AI in democratic participation: from using language models to simulate human perspectives, to designing AI as a thought partner in urban planning, to developing ranking systems for large-scale participatory budgeting. Across these applications, the question remains: when we align AI, for whom?

Tobias Henschen: Algorithmic decisions and equality of opportunity

My paper models algorithmic decisions as optimizing payoff functions that are subject to a constraint of algorithmic fairness. It argues that conditional statistical parity (CSP) is the required constraint because the ideal of equality of opportunity (EOP) seems to be violated whenever we think that algorithmic decisions are unfair, because CSP requires the equality of

decision rates across groups, given a set of “legitimate” variables, and because a general theory of EOP (specifying conditions under which competing conceptions of EOP morally constrain decision-making) can be shown to determine the variables to be included in this set. My paper also argues that the optimization of payoff functions constrained by CSP involves normative judgments, and that these judgments bias algorithmic decisions, while not necessarily violating race or gender neutrality.

Sune Holm: On the use of decision thresholds to convert algorithmic probability scores into decisions

On the dominant statistical approach, algorithmic fairness is a matter of setting appropriate score thresholds. In this article I criticize this approach. I first argue that fairness is essentially an individualistic concern, which is not taken seriously when fixed thresholds are applied. Then I present a method for operationalizing fair treatment of individuals using a score-weighted lottery when converting their scores to decisions. And, finally, I respond to some objections to my proposal and outline some unresolved challenges.

Vincent Müller: AI Ethics: Universal Scope and Method

This paper detects symptoms of a problem in AI ethics, namely that while (1) there is a significant societal demand for AI ethics in the last 5 years or so, (2) there is insufficient supply of expertise and results in AI ethics. Instead, there is a diversity of “experts” of AI ethics who disagree on what the problems of AI ethics are, which ones are important, how one works on these, who should work on these, and what the expected outcomes are. The problem with (1) and (2) is that we do not know what the scope of the field is, and what its methods are. – The paper proposes a diagnosis why these problems occur, namely (3) that the scope of AI ethics is nearly universal, because AI is about intelligence, and (4) that the methods of AI ethics have to recognise that universality, as well as the different aims that are pursued in the field, and the many uses in socio-technical systems that exist or will exist. Finally, we propose a therapy for unclear scope and method through a two-dimensional matrix of ethical themes vs. socio-technical uses of AI. This method allows us to locate canonical problems, positions and arguments for the field; at the same time, it provides space for a multitude of approaches, for future developments, and for diversity of expertise.

Hadeel Naeem: Situating the epistemic agent in the AI age

Philosophers and epistemologists of AI often focus on how AI systems harm users, and when doing that, they usually look for something that has gone wrong. For instance, they look at biased or narrowing output, inquiry that is blocked, agency that is diminished or distorted, or skills that are lost. When studying harms that specifically befall non-dominantly situated agents, many researchers describe how similar harms affect, or often disproportionately affect, such agents. Even though many of these accounts speak of epistemic agents, they treat agency mainly as something AI systems affect. As a result, the non-dominantly situated agents often sound as if they are merely passive victims of AI harms. What this focus misses is when epistemic agents, especially the non-dominantly situated agents, exercise epistemic agency to get things right or responsibly form true beliefs with AI. When we look at how two differently situated epistemic agents exercise epistemic agency to get things right, we see that the non-dominantly situated agents often have to put in far more epistemic labour than the dominantly situated agent to achieve the same output — that is, a responsibly formed true belief.

Katia Parshina: Philosophical accounts of bias and inductive bias

When we talk about bias, we normally mean some systematic error that compromises objective judgment; in machine learning, we often focus on mitigating algorithmic bias, i.e., unwanted model behavior. At the same time, we know it is necessary to introduce some inductive assumptions about the learning problem we want our algorithm to solve – that is, some **inductive bias**. It is clear why the adjective ‘inductive’ is there – the algorithm uses the observed data to classify the data yet unobserved, so the type of inference at play is indeed

inductive. The noun 'bias', however, seems very out of place. It makes 'inductive bias' sound almost like an oxymoron – as if it leads us to knowledge by compromising the very judgments through which we acquire it. Let's see whether the philosophical accounts of bias do a good job at untangling this semantic nonsense.

Eric Raidl: Readings without a measurand: why ML models are not instruments

Machine learning models are increasingly described as scientific instruments: computational microscopes, stochastic measuring devices, soft sensors, indices. I distinguish two claims behind such talk. The artifact thesis says that models are tools we use for epistemic purposes. The reading thesis says that their outputs are measurement outcomes. Drawing on metrology and the philosophy of measurement, I argue that the reading thesis fails. An instrument's indication becomes an outcome by an inference licensed by a measurement principle or a model of the measurement process, and calibration against standards with stated uncertainty secures that inference. Across five kinds of instruments (observational, physical, model-based, protocol, index), ML outputs lack such an inference: they are indications without outcomes, projections rather than extractions, learned relations rather than known principles, with no theory of when uniformity breaks. ML's vocabulary of "ground truth", "calibrated", "validated" and "benchmark" borrows metrology's words without its traceability chains. This matters for values in science. Metrology institutionalizes the division Jeffrey defended against Rudner: scientists assign probabilities, and decision-makers accept against tolerance limits. That division presupposes an outcome with stated uncertainty. An ML score becomes a probability only after probabilistic recalibration, and using it requires three acceptances: projecting to the case, assuming the case is drawn from the training distribution, and setting a threshold. The assign/accept division sees only the last. Value judgments therefore enter before any decision threshold is set.

Gerhard Schurz: Value-neutrality and value-engagement in science and machine learning

This talk develops and defends a new version of the thesis of scientific value neutrality.

The first part defines the value-neutrality thesis more precisely: The justification of scientific knowledge should be independent of fundamental non-epistemic value assumptions. Although extra-epistemic interests may influence the choice of research question, I argue that once the research question is fixed, it screens off such interests from the justification of epistemic methods.

The second part shows how the value-independence of scientific knowledge can be reconciled with scientific value engagement through the method of hypothetical value statements. These are conditional value recommendations derived by means–end reasoning from hypothetically assumed values supplied not by scientists but by users of science. I explore the implications of this approach for machine learning.

The third part argues that the same hypothetical method can address the problem of "inductive risk" without recourse to categorical value judgments. Risks of error should be explicitly stated and made transparent to non-experts through hypothetical cost–benefit analysis.

The outlook applies these ideas to further controversial issues, including the "new demarcation problem".