BOVAY ENGINEERING AND APPLIED ETHICS WORKSHOP

AI VALUE ALIGNMENT

Time	Speaker	Title	Institution
8:30 – 9:00		*Light Breakfast*	
9:00 – 9:50	Andrew Burnside	Don't Be an AI Hypochondriac (Comments by Martin Peterson)	New Jersey Institute of Technology
9:55 – 10:45	Ava Thomas Wright	*Humanity* Compatible (Comments by David Koepsell)	Cal Poly
10:45 – 11:00		*Coffee Break*	
11:00 – 11:50	Samantha Bennett	Testing the Ropes: A Dynamic Epistemic Logic Approach to AI Safety (Comments by Mitchell Roberts)	Stanford University
11:55 – 12:50	Pamela Robinson	Uncertainty-Sensitive Oughts: Lessons for AI Alignment (Comments by Glen Miller)	University of British Columbia
12:50 – 1:40		*Lunch Break*	
1:40 – 2:35	Arianna Manzini	T.B.A. (Comments by Martin Peterson)	Google DeepMind
2:40 – 3:30	Michael Anderson	AI as an Ethical Collaborator: Moving Beyond Value Alignment to Principled Ethical Reasoning (Comments by Erich Riesen)	University of Hartford
3:30 – 3:45		*Coffee Break*	
3:45 – 4:35	Erich Riesen	Expert Voting: A Better Metanormative Approach (Comments by Brandon Wadlington)	Texas A&M University
4:40 – 5:35	Rebecca Raper	A Pragmatic Take on the Problem of Many Values in AI Alignment Research (Comments by Dylan Shell)	Cranfield University
5:35 – 6:00		*Break*	
6:00 – 7:15	Arianna Manzini	**Bovay Lecture: The Ethics of Advanced AI Assistants**	Google DeepMind

The workshop takes place on the campus of Texas A&M University, **Thursday, April 10th**. The Bovay Lecture will be hosted in Heldenfels Hall (HELD), room 100. All other presentations will take place in the YMCA building, room 401.

TEXAS A&M UNIVERSITY
DEPARTMENT OF PHILOSOPHY

Funded by the Bovay Foundation. Organized by Martin Peterson and Erich Riesen. For questions, please contact Erich Riesen (emriesen@tamu.edu). An updated program will be provided one week before the event.

Andrew Burnside

New Jersey Institute of Technology

Don't Be an AI Hypochondriac: Much recent work in the value theory of autonomous and intelligent systems (AIS) circulates around two issues. First is the alignment problem: the problem of producing AIS whose values align with humanity's interests. Second, superintelligence: the potential for AIS to develop intelligence which would surpass even the most intelligent humans. An increasing number of authors suggest that the concatenation of these problems should direct interest to the long-term potential for misaligned, superintelligent AIS. They argue for a policy stance which we describe as "hard alignment", i.e., cooperating with technological experts to avoid hypothetical scenarios where AIS disempower humanity. On the other hand, we describe our view as "soft alignment." Considering the lack of adequate evidence for hard alignment's radical claims, the finite resources and attention of policymakers and experts are best served by devoting, at best, a modest amount of time, attention, and resources to policies that manage the moral risk involved in misaligned AIS. We argue for the adoption of policies which manage the everyday risk involved in misaligned AIS rather than long-term existential risks, which are difficult to quantify.

Ava Thomas Wright

Cal Poly

***Humanity* Compatible:** In this paper, I will argue that human-compatible AI (HCAI) agents should interpret human behavior as efforts to act autonomously in the Kantian sense of that term. Human rational agency is not limited to finding instrumentally efficient means to maximizing the satisfaction of preferences presumably given by natural animal instincts or social forces external to us. Unlike other animals (or, indeed, machines), we are free to assess, and then accept or reject, the reasons our various incentives may propose for acting. Kant refers to this capacity for autonomy as our "humanity," and it is the foundation of our rights and responsibilities. I thus argue for humanity compatible AI. The Kantian HCAI would continuously align its behavior to help us realize our autonomous choices. At the same time, it would respect our rights to freely choose our own ends, so long as we did not wrong others. The Kantian HCAI would thus never "assist" by means of coercion, deception, or manipulation because these means are incompatible with helping us to act autonomously.

Samantha Bennett

Stanford University

Testing the Ropes: A Dynamic Epistemic Logic Approach to AI Safety: In 'AI Safety: A Climb to Armageddon?', Cappelen et al. (2024) present an argument against artificial intelligence (AI) safety measures, with the counterintuitive conclusion that implementing safety measures leads to worse outcomes upon eventual AI system failure. While their argument is attractive in virtue of its simple utility framework, this paper argues that it relies on a fundamentally flawed analogy—the Doomed Rock Climber—which fails to capture crucial distinctions between different types of AI safety measures. I argue that Cappelen et al.'s argument applies only to what Lazar (2024) calls "AI companions"—personal-use AI systems already deployed to end users—while neglecting the broader landscape of AI development. Instead, I suggest that increased safety testing and auditing during development phases, when properly contained and controlled, lead to demonstrably safer technologies upon deployment.

Pamela Robinson

University of British Columbia

Uncertainty-Sensitive Oughts: Lessons for AI Alignment: A primary task of normative theorists is the analysis of concepts that are sensitive to uncertainty and ignorance. We try to determine what makes beliefs justified given our evidence, what makes actions morally permissible given our uncertainty about their effects, and so on. I will call this general project that of finding uncertainty-sensitive oughts. This uncertainty-sensitive project can be seen as part of its own larger alignment project, in which we aim to align ourselves with valuable normative principles. We want uncertainty-sensitive oughts because we want to know what to do in the real circumstances we find ourselves in, where our uncertainty and ignorance can't be idealized away. I argue that we can learn some lessons for AI alignment from the uncertainty-sensitive project, including: that there are unavoidable trade-offs between value and alignment, that there is a higher-order problem that frustrates attempts to reach certainty about alignment, and that there is reason to want AGI (or powerful and autonomous AI agents) to have at least the same capacity for uncertainty as we have.

Arianna Manzini

(T.B.A.)

Google DeepMind

Michael Anderson

University of Hartford

AI as an Ethical Collaborator – Moving Beyond Value Alignment to Principled Ethical Reasoning: The problem of value alignment, ensuring AI systems act in accordance with human values, has been widely acknowledged as a core challenge in AI safety. However, existing approaches often rely on aggregating human preferences, whether through surveys, reinforcement learning from human feedback (RLHF), or other participatory mechanisms. These methods risk conflating what is preferred with what is morally justifiable, sidestepping the deeper issue: whose values should AI be aligned with, and how should those values be determined? Our work proposes a crucial shift: rather than treating AI as a passive system that merely aligns with human preferences, we argue that AI should be designed as an active ethical collaborator, capable of engaging in principled ethical reasoning alongside human experts. We leverage AI's ability to analyze vast datasets of ethical discourse, spanning historical, cultural, and philosophical perspectives, to synthesize moral principles that are not merely based on contemporary consensus but reflect the most rigorous moral reasoning available.

Erich Riesen

Texas A&M University

Expert Voting – A Better Metanormative Approach: Societies often face decisions under moral uncertainty. AI provides an especially salient example. How should AI agents behave when confronting difficult moral choices? Metanormative theories suggest that we handle moral uncertainty on analogy with decision theoretic treatments of empirical uncertainty. For instance, Bogosian (2017) suggests that "moral machines" ought to maximize expected moral value by weighting the value of each alternative by the probability that the moral theory assigning it is correct. In this presentation, I offer a better metanormative approach. I argue that we should treat value alignment as a social choice problem, but one in which the preferences of experts (and not the public) are aggregated. Expert voting, unlike public-compromise, does not illicitly derive an ought from an is. Moreover, it avoids key problems inherent in the application of traditional dominance or expected moral value approaches to practical ethics.

Rebecca Raper

Cranfield University

A Pragmatic Take on the Problem of Many Values in AI Research: One of the central problems in Artificial Intelligence (AI) Alignment Research is what might be termed the problem of many values: if we are looking to embed ethical principles into AI decision-making, whose principles do we embed? In other words: how do we capture the diversity of human moral opinions (across cultures, individuals...) in the decision architecture of an AI system, so that its behaviours are aligned with 'good'? This is a problem plaguing computer scientists, who want to assure the good behaviour of their machines. A seemingly obvious solution is to apply democratic principles, in other words, a 'vote', on what behaviours get embedded into the AI's architecture, but this approach is problematic because it can seem to dilute moral opinion that sits outside the normal range. So, how do we capture different moral opinions? If we take a traditional Pragmatic stance on morality, that there is no absolute morality we know, but that as humans we are driving a path towards absolute morality, we start to see that attempting to distil moral opinions into a machine is wrong-headed because, it assumes that those moral opinions are absolute. So where does this leave us? Instead, we need to embrace the diversity of opinion, looking not to capture moral thoughts, embedding these in the machine, but replicate the processes by which they come about. We are left with a new pursuit: to understand the cognitive basis for moral decision-making, rather than trying to consolidate every moral opinion there might ever be.

The Bovay Foundation, in conjunction with the Department of Philosophy, presents:

The Bovay Lecture

Arianna Manzini

(Google DeepMind)

The Ethics of Advanced AI Assistants

Thursday, April 10th | 6:00-7:15 | Held 100

Open to the public.
Organized by Martin Peterson and Erich Riesen.
For questions, please contact Erich Riesen (emriesen@tamu.edu).