Navigating Epistemic Gaps and Ethical Risks in AI Alignment: A Pluralistic Approach to Human Values

Simone Zhenting Mao, Harvard University

International Conference on Large-Scale AI Risks: Control, Governance, and Ethics
Ramsey Lab & Chair Ethics and AI, Institute of Philosophy, KU Leuven
26-28th May 2025, Belgium
https://www.kuleuven.be/ethics-kuleuven/chair-ai/conference-ai-risks

Navigating Epistemic Gaps and Ethical Risks in AI Alignment: A Pluralistic Approach to Human Values

Good afternoon, everyone.

I’ve never been great at self-introduction, so I prefer to let my research speak on my behalf.

As artificial intelligence systems grow more powerful, their ability to act autonomously and at scale has introduced a profound and urgent challenge: Value alignment. The alignment issue concerns whether AI can accurately understand and implement human moral intentions and how to identify suitable ethical guidelines in cross-cultural, cross-value contexts. How do we ensure that these systems align with the deeply contested, often implicit, and morally complex spectrum of human values? We face the risks of *misalignment*, where AI pursues goals that diverge from our own; *value conflict*, where cross-cultural ethical systems clash; and *capability exceedance*, where AI surpasses human control, leading to unpredictable actions.

All of these are exacerbated by a central tension: we have no consensus on what constitutes “human values,” let alone any actionable standard for aligning AI with them. I’m a political theorist. What concerns me is this philosophical instability—we lack a coherent theory of what “human values” even are, and how can we hope to align anything to them?

This reveals a foundational issue—what I call the *epistemic gap*—the gap between our limited ability to define moral goals and the machine’s powerful capacity to optimize whatever we give it. In short, there are two sides: Learnability: Can values even be learned by machines if we ourselves cannot agree on them? Representability: Can they be expressed in forms that machines can understand and act upon? This gap grows more dangerous in a world marked by cultural pluralism. It risks becoming a project of selective representation—who decides what counts as a “human value,” and on what terms? In this context, AI alignment is not just a technical challenge, but a philosophical and political one.

Today, I will argue for an approach grounded in pluralism—drawing on the work of Isaiah Berlin, John Rawls, Charles Taylor, and others. This means building frameworks that allow diverse values to coexist. I’ll close by briefly discussing the global governance implications of this pluralist approach.

Many current alignment approaches adopt the idea that there is a single, universal value function that can guide all decisions. Utilitarianism, in particular, dominates the discourse, positing that any conflict can be resolved by choosing the option that maximizes *utility*—be it happiness, preference satisfaction, or well-being. This view represents an ethical *monism*—rooted in the search for a unified theory of the good. According to monists, fundamental value conflicts can be resolved by appealing to a single formula for ranking or trade-offs that applies in every case. Liberty takes precedence over equality when doing so maximizes utility, and vice versa.

So, Berlin’s first insight lies in his recognition of a fundamental truth about human society—not merely a matter of appearances, but of the deep structure of our ethical experience: the underlying divergence of human values is inherent and unavoidable. It must be acknowledged as incommensurable—there is no simple formula to resolve it.

This insight is not unique to Berlin. Thinkers like Nietzsche, Vico, and Herder also highlighted its significance. Then Deleuze, Rawls, Leo Strauss—and many more could be listed. But I want to push further: most past philosophical systems are monistic, arguably nearly all of them, at least in the Western tradition. I argue this is not incidental, but intrinsic to the structure of philosophy itself. It is a defining characteristic of philosophy, both a product and a demand of the history of philosophy.

But Berlin sharpens the point: value conflict is not an anomaly; it is the norm. It is not something to be “solved” by identifying a master value—it is a reality to be respected.
(1) Is value pluralism real? Yes.
(2) Does this imply relativism? Not necessarily—pluralism is not the claim that “anything goes,” but that *multiple, incompatible yet legitimate values can coexist*.
(3) What happens when plural values clash? In some cases, no neutral principle can decide; we must make hard, context-sensitive choices.
(4) What are the political implications? Any governance or alignment strategy that ignores pluralism risks domination—imposing one moral framework over others.

Value pluralism is not just an empirical reality in a world we live in—it is a normative foundation for coexistence. Berlin pointed out that the essence of politics is not the pursuit of truth, but the management of tensions between incompatible goods. Accordingly, AI alignment requires institutional safeguards. It calls for “balancing ethic” rather than a “maximizing ethic” —recognizing that every choice may entail the sacrifice of other legitimate goods.

Let’s talk about a related question. If a globally adopted AI leads to epistemic and cultural homogenization, imagine the consequences: everyone thinking, expressing intentions, interests—even emotions—through the same framework of values and reasoning. The erosion of epistemic and cultural diversity through this form of digital colonization is deeply troubling.

Imagine if different civilizations, nations, and cultures could create their own culturally grounded AIs—especially generative AI. But here lies a paradox: when people use AI systems built around specific value frameworks, those values get reinforced over time. In nation-states or authoritarian regimes, and even beyond, it risks narrowing the range of thought.

Will this suppress more diverse, dissenting voices? Will it deepen the tension between tribalism and globalization? And more fundamentally—are we approaching a new kind of cultural or ideological warfare, where AIs become instruments in battles over values, identities, and worldviews? AI learns from humans—it reflects human society and our value choices. We need to find a way to prevent AI systems from simply becoming a new arena for these dynamics or a battleground for existing conflicts. The challenge is to find a framework to regulate and govern them, and avoiding the imposition of values is crucial.

But here, there are more questions than answers. So I invite the audience to reflect together:

One approach Rawls offers is overlapping consensus—where individuals with different moral doctrines converge on shared principles of political justice. His theory builds on the tradition of the social contract, particularly Rousseau’s notion of the *general will*. Later, I will explore an alternative that emerges from this tradition, if time permits.

His starting point arises from a similar view of the modern world to Isaiah Berlin’s.

In today’s world of value pluralism, cultural diversity, and conflicts of interest, AI systems are expected to align with “human values.” But human values are dispersed across cultures, religions, and worldviews. This reasonable pluralism, as Rawls describes, is a defining feature of modern global society. Different countries, institutions, and individuals often hold conflicting views on what counts as "good" or "just".

Expecting AI to reflect a single, comprehensive moral view is unsustainable and normatively dangerous. Using overlapping consensus as a foundation has some advantages: to be trusted globally. We need a consensus mechanism that can accommodate diverse values—not grounded in any one metaphysical or ethical system, but in principles of public reason and institutional legitimacy. The value foundation of AI should rest on shared, cross-cultural ethical minimums—widely accepted across diverse moral communities.

This legitimacy cannot be a functional compromise to avoid conflict—but must instead reflect moral sincerity. This commitment imposes three normative design requirements on AI systems: explainability, predictability, auditable reasoning.

I know it's hard to achieve. True alignment must be grounded in publicly justifiable reasons. But this assumes institutional neutrality, both in liberal states and in AI governance. In reality, institutions often reflect the interests of the advantaged. We must ask: who defines it and under what conditions? This demands the inclusion of multiculturalist perspectives that question how value systems are constructed, whose values are prioritized, and which voices are excluded.

AI alignment faces the very challenges faced by modern liberal societies. It requires an approach to address:

1. The legitimacy crisis in pluralist societies;
2. The need for stable coexistence through public reason;
3. The normative goal to find common ground without erasing differences.
4. It calls for a new conception of political justice—distinct from comprehensive moral doctrines.

These are the core concerns of my work in political theory—impacting broader topics than AI alignment. My effort is to offer philosophical tools to build a widely acceptable ethical framework. My recent research has developed a Justice-Oriented Ethical Impact Assessment for AI that brings *justice as difference* into AI governance. This model aims to institutionalize cultural pluralism within global AI governance.[1]

So we move to the third part: The politics of recognition helps us move beyond "value uniformity" toward a richer framework, which centers difference, cultural respect, and equitable alignment across diverse groups.

The very simple idea at the core of the politics of recognition is that, our identities—our sense of who we are as individuals and as members of a particular community—are of tremendous value and importance to us, and for this reason are deserving of recognition and respect. Recognition is understood to be partially constitutive of identity, our identities are partly shaped by their recognition or non-recognition by others. When others recognize the importance and worth of our

[1] Mao, S. Z. (2025). Justice-Oriented Ethical Impact Assessment for AI: Institutionalizing Cultural Diversity in Global AI Governance. The 3rd UNESCO Global Forum on the Ethics of AI, Bangkok, Thailand.

identity, this contributes significantly to our sense of personal security, self-respect and well-being—our feeling that we are regarded as equal, and equally valued, members of society. Conversely, being unrecognized—or recognized in a demeaning way—can result in identity-based oppression and psychological harm. Thus, recognition is not a courtesy, but a basic human need grounded in our status as moral equals.

It offers important insights for AI ethics by revealing the cultural and structural biases behind so-called "neutrality." Current AI systems widely adopt datasets, value definitions, and design standards shaped by majority cultural contexts. Just as national laws often adopt the language and institutions of the majority as 'neutral,' AI systems can unrecognize and unintentionally oppress minority groups.

Many seemingly "consensual" AI ethical principles—such as transparency, privacy, and accountability—are often defined by a small number of corporations, Western institutions, or elites. For minority cultures to be "respected," they must first conform to majority standards.

As critics have pointed out, when recognition can only occur within a value framework set by the majority, minority groups can only be "recognized" passively, rather than actively claiming recognition on their own terms (Bannerji 2003). In response, Taylor argues that anyone engaged in a process of cross-cultural evaluation inevitably begins this process from within their own contingent set of moral standards—what Taylor calls their own moral horizons. The important question is, where are they prepared to go from there? The non-ethnocentric path lies in intercultural dialogue— approaching others with an openness to revising one's own moral horizons in the process.

Thus, structural correction, conducting cultural bias audits on the default data and behavioral models is important to identify underlying structural inequalities. Who sets the rules? Who speaks for humanity? Which groups, cultures, and beliefs have the right to participate—and who is excluded? A multi-stakeholder negotiation mechanism must be established. Alignment is not only about embedding the preferences of decision-makers; it also involves the identity-conscious design of data structures and algorithms.

Tech developers are not moral legislators. Processes of value recognition should be integrated into the entire AI lifecycle, through pluralistic governance to ensure cross-cultural representation and legitimacy in AI.

From its inception, AI has possessed seven forms of internationality:

1. The global collaboration in its epistemic origins;
2. The transnational spread, development, and deployment of technology;
3. The global nature of ethical and value conflicts;
4. The cross-border nature of risks and responsibilities;
5. The multipolar trend in governance;
6. The global inequality in data resources and linguistic ecosystems;
7. The necessity of cooperative governance and overlapping consensus.

These realities demand building a democratic legitimacy mechanism for AI alignment—such as international AI treaties, transnational ethical councils, and multilingual public forums.

AI is not a “national or regional technology,” but a form of global intellectual engineering. Its legitimacy must come not from central command, but from distributed consent—from systems that listen as much as they optimize, and from governance that reflects plural perspectives.