Views on AI Aren't Binary — They're Plural (Extended Abstract)

Thorin Bristow[1], **Luke Thorburn**[2], **Diana Acosta-Navas**[3]
[1]Independent Researcher
[2]King's College London
[3]Loyola University Chicago
mail@thorinbristow.com, luke.thorburn@kcl.ac.uk, dacostanavas@luc.edu

In March 2023, an organization called the Future of Life Institute published an open letter calling for all AI labs to "pause for at least 6 months" the training of frontier AI systems due to concerns about "potentially catastrophic effects on society." Some of the most deeply felt criticisms of this letter came from a group of researchers who have also been working toward mitigating the societal risks associated with AI. They argued that the letter placed disproportionate attention on speculative risks, disregarding harms that the AI industry is already causing. This exchange brought broader attention to tensions between two overlapping communities, "AI Ethics" and "AI Safety."

While the overall debate has evolved and new narratives have emerged, the perceived binary conflict continues to influence how debates over AI are framed, and the extent to which people working on AI development and governance can trust each other and collaborate. Our aim for this paper is to support those navigating the politics of AI by providing context and language with which to understand these two perspectives. While we focus on "AI Ethics" and "AI Safety," the general lessons apply to other related tensions, including those between accelerationist ("e/acc") and cautious stances on AI development.

The first section documents the false binary. Drawing heavily on grey literature including mainstream press, personal blogs, and social media posts, we describe the (false) stereotype of each group, summarize the complex grievances between the two communities, and the reasons why the language commonly used to describe them is fraught. The second section argues that the binary, as it is commonly presented, is in many ways inaccurate. We articulate points of commonality between the two communities, including a widespread basis of goodwill. We draw on empirical evidence to argue that the space of perspectives on AI is more diverse than can be accurately modelled as a binary, and note that narratives about the risk of powerful, difficult-to-control AI systems, regardless of their truth status, can be co-opted by corporations for financial gain. Finally, in the third section, we propose five concrete strategies for managing these tensions and moving forward constructively:

1. **Build holistic institutions.** Differences over priorities can be resolved organizationally, with different teams or departments assigned responsibility for different concerns.
2. **Conduct broad church collaborations.** Bringing together diverse stakeholders from various disciplines and backgrounds can help foster a more comprehensive understanding of AI's impacts and develop inclusive solutions that address the full spectrum of concerns.
3. **Where possible, test contentious claims empirically.** Many contested claims can be evaluated empirically, and producing empirical answers can create a source of common ground.
4. **Be a surprising validator.** When people act as surprising validators — calling out when they agree with the "other side" — this helps subvert the false binary and avoid pluralistic ignorance (where everyone is afraid to speak up because they feel they are the only person in their community who thinks differently).
5. **Where appropriate, develop processes for collective input.** In some cases, new processes for collective or democratic input and oversight (e.g., of product design decisions and fine-tuning policies) might help ensure AI systems are responsive to diverse perspectives.

Recent policy debates over climate change and the SARS-CoV-2 pandemic demonstrate how division and politicization over societal challenges can undermine our collective response. The shoehorning of opinions into mutually exclusive groups, sometimes called partisan sorting or ideological sorting, has long been recognized by political scientists and conflict scholars as a risk factor for the escalation of destructive conflict. High levels of sorting make it easier to stereotype and pigeonhole people and reduce representation of nuanced, cross-cutting positions. Conversely, low levels of sorting increase the number of "surprising validators" who help make groups legible to one another, and increase the likelihood that any majority will include representatives of any minority, reducing the risk of majoritarian tyranny. Avoiding sorted, us-vs-them conflict — *high conflict* — in the broad community of people working on AI development and governance will help us ensure that the impacts of this emerging technology are inclusively beneficial.

For the full paper and references, see:

https://bit.ly/ai-false-binary