In the first few decades of the internet the adage was, “Don’t believe everything you see on your screen.” But now virtually everyone believes exactly whatever appears on the screen. This notion of belief gets even more troublesome as Artificial Intelligence (AI) begins to become more prevalent in technological application. This has caused AI to quickly become intertwined into our daily life, going from virtually none to constant. When you have a question instead of googling, you use an AI product. When you can’t figure something out, an AI chatbot has the answer you desire. When you need help predicting an outcome, AI will do it for you. With so much trust put on AI to produce correct answers a question must be asked, how should we align the answers that AI produces to our values?

The most worrying potential is if an AI is misaligned to our core human values. For instance, what happens when large language models start to spread misinformation? A recent tragedy comes to mind: a teenager took their own life after an AI chatbot pushed them to do so(Associated Press, 2024). This highlights, on the most extreme end, what the consequences are when AI is not aligned with our values.

The first step to aligning AI is to establish a framework. For this a constitutional alignment can be initially used in combination with pluralistic ideals. Pluralism acknowledges that we don’t need to nor do we have everything figured out at this moment. A pluralistic approach recognizing that we can build our AI framework gradually, continuously aligning it to our values as they change. The constitutional alignment offers a baseline alignment of AI with broader legal principles to help to create a foundational level of value alignment. For instance, if we believe in the inherent equality of all individuals and the prohibition of discrimination, our AI algorithms must reflect this belief. If AI is used in employment decisions, it must comply with anti-discrimination laws. By establishing these basic principles, we can avoid contentious debates over implementation, as these issues are already addressed within our legal framework. Constitutional approach bridges the gap between the unaligned, or even misaligned, models that are present right now and a future more aligned model.

To continue the alignment after an initial constitutional alignment we can follow Raphael Milliere who breaks the alignment problem into two things, “(a) identifying fair principles to guide the behaviour of LLMs that can be endorsed despite reasonable pluralism in beliefs about social and moral norms; and (b) finding effective strategies to steer the behaviour of LLMs in accordance with these guiding principles.”(Milliere, 2023) This process takes significantly more time than a constitutional alignment. The second part is somewhat the easier path forward, as most of it can be solved through engineering principles. The first part identifying fair principles and agreeing to them will take the most time.

The alignment of LLMs with fair principles is a complex and time-consuming process, primarily due to the need for broad consensus on ethical standards amid diverse beliefs. Ultimately, the goal is not just to create effective AI applications, but to develop systems that resonate with a shared sense of fairness and justice, fostering trust and acceptance in their use.

Associated Press. (2024, October 25). *An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges*. https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0

Milliere, R. (2023, November 3). *The alignment problem in context*. Department of Philosophy, Macquarie University. https://arxiv.org/pdf/2311.02147