Contribution ID: **27**

Type: **not specified**

AI Alignment: Problem of Diversity & Physics

Tuesday, 8 October 2024 10:55 (25 minutes)

The AI Alignment Problem involves aligning AI behavior with human intentions, addressing both technical and ethical concerns. In the first part of my talk, I exemplify the AI Alignment problem in the context of gender inequality in GPT4o. The AI Alignment problem effects also particle physics, where AI is essential for tasks like event tagging or event generation, and has the potential to enable frontier precision predictions in the standard model. Therefore, ensuring reliability of AI systems is crucial for the future development of particle phyiscs.

To avoid unaligned behavior, benchmarks are typically used to assess the performance of AI models. However, benchmarks are insufficient to assure aligned AI systems. A particular failure mode is strategic underperformance on benchmarks, leading to misaligned behavior of the AI during deployment. In the second part of my talk, I present a novel experiment inspired by deep learning theories rooted in physics to detect this type of underperformance.

Presenter: KREER, Phillip

Session Classification: Short talks