

Spark Talk

Jacob Tsimmerman

AI Safety and Mathematics

What does it mean for an AI system to be safe? There are of course deep questions and complicated coordination problems regarding (cyber)security and policy. But even if we get that right, it turns out to be extremely complicated to design and direct a supertelligent system such that it acts in ways we would consider "safe" or "good," or even just to avoid catastrophic consequences. The plan is to chat about this problem, why it's hard, and why mathematics might play a big role in attempting to solve it.