AI Safety

September 30, 2025

Raj Rajagopalan
Internautica, Inc.
ntuple.com

What is AI Safety ?

A field focused on ensuring artificial intelligence systems operate reliably, ethically and beneficially, aligning with human values and preventing unintended harm.

How is it different from AI Security ?

Safety is focused on Risk while Security is about Threats.

As threats are a subset of risks, Security is a subset of Safety.

Where do these risks originate ?

AI Risk come from four major sources:

- Malicious Use
  - Deliberate use of AI to create harm
- AI Race
  - Competitive pressures leading to unsafe practices and outcomes
- Organizational Risks
  - Inadvertent issues originating from the complexity of AI and the capabilities of organizations deploying them.
- Rogue AI
  - Problems from the creation of intelligent systems that operate in unintelligible ways

Macro Risks

- AI Doom
- Totalitarian AI Dominance
- Plutocratic Hegemony
- Privilege Polarity
- Humanity’s lack of Purpose

AI Doom

- AI 2027
  - Project calculating the ongoing probability of AI killing us all in 2 years
- Peter Thiel and the Antichrist (WSJ)
- AI Restraint and Pause open letter

Totalitarian AI Dominance

“Condoleezza Rice consistently emphasizes that the development of AI is a geopolitical competition between democracies and authoritarian regimes.

She warns that if the U.S. and its allies do not lead in this field, they risk ceding control of this transformative technologies like China, which may use it in ways that threaten democratic values and international security.

She stresses the importance of winning this race to ensure that the guardrails of democracy and ethical considerations are built into the technology.”

Plutocratic Hegemony

Large corporations and their billionaire owners will monopolize the benefits of AI as only they have the resources to develop AI capable of powering the economy.

Privilege Polarity

AI's productivity gains estimated at 10x the Industrial Revolution by Demis Hassabis, exacerbate societal inequalities much as they did in the previous era but at a significantly elevated scale.

Humanity loses Purpose

If AI can do all things that humans can but only better, will people feel angst at an existential level ?

John Maynard Keynes believed increased productivity would eventually allow for a 15-hour workweek, leading to greater leisure time, which people would use to focus on "wise and agreeable and well" living, shifting focus from money to non-monetary satisfactions and personal growth once basic needs were met … but this has not happened.

Organizational Risks

- Data leakage
- Legal risks
- Reputational risks
- Process manipulation

How do we address these risks

- Visibility
  - CoT
  - Monitoring
  - Adversarial Testing
- Safety Engineering
  - Risk Composition and Reliability Measurement
  - Safe Design Principles
  - Application of Accident Models from other fields
- Beneficial AI and Machine Ethics
  - Law
  - Economic incentives
- Governance
  - Standards and Guidelines
  - Liability
  - Constraints on Compute, Data and Algorithms

And yet …

- Behavior of Complex Systems
- Collective Action Problem
- AI’s rapid evolution and lack of scientific understanding