COURSE CERTIFICATE

2024 | AI Safety Alignment

AI Safety Alignment

COMPLETED BY

Laxmi Dhital

02.05, 2024

The AI Safety Alignment course explains why it is plausible to see the emergence of human-level AI, the potential risks associated, and what solutions are currently being explored to mitigate these risks. Participants also dives into the specific technical challenges of AI alignment, such as the problem of reward misspecification - exploring concepts like optimization, Goodhart's Law, and different approaches to learning human preferences - and goal misgeneralization - examining concepts like mesa-optimization, inner alignment, and deceptive alignment - aiming to provide a deep understanding of the main technical challenges in AI alignment.

The course textbook was created by Charbel-Raphael Segerie, Markov Grey, Jeanne Salle and Vincent Corruble. The course is organized by the AI Safety Bergen in collaboration with AI Safety Collab.

Christian Kårbø Engelsen
AI Safety Bergen
Course Facilitator

Anette Molvik Nilsen
AI Safety Collab
Project Lead

Certificate issued July 13, 2024 | Certificate verification: AISafetySweden.com/verify/ | Verification code: LTMqt062Bv