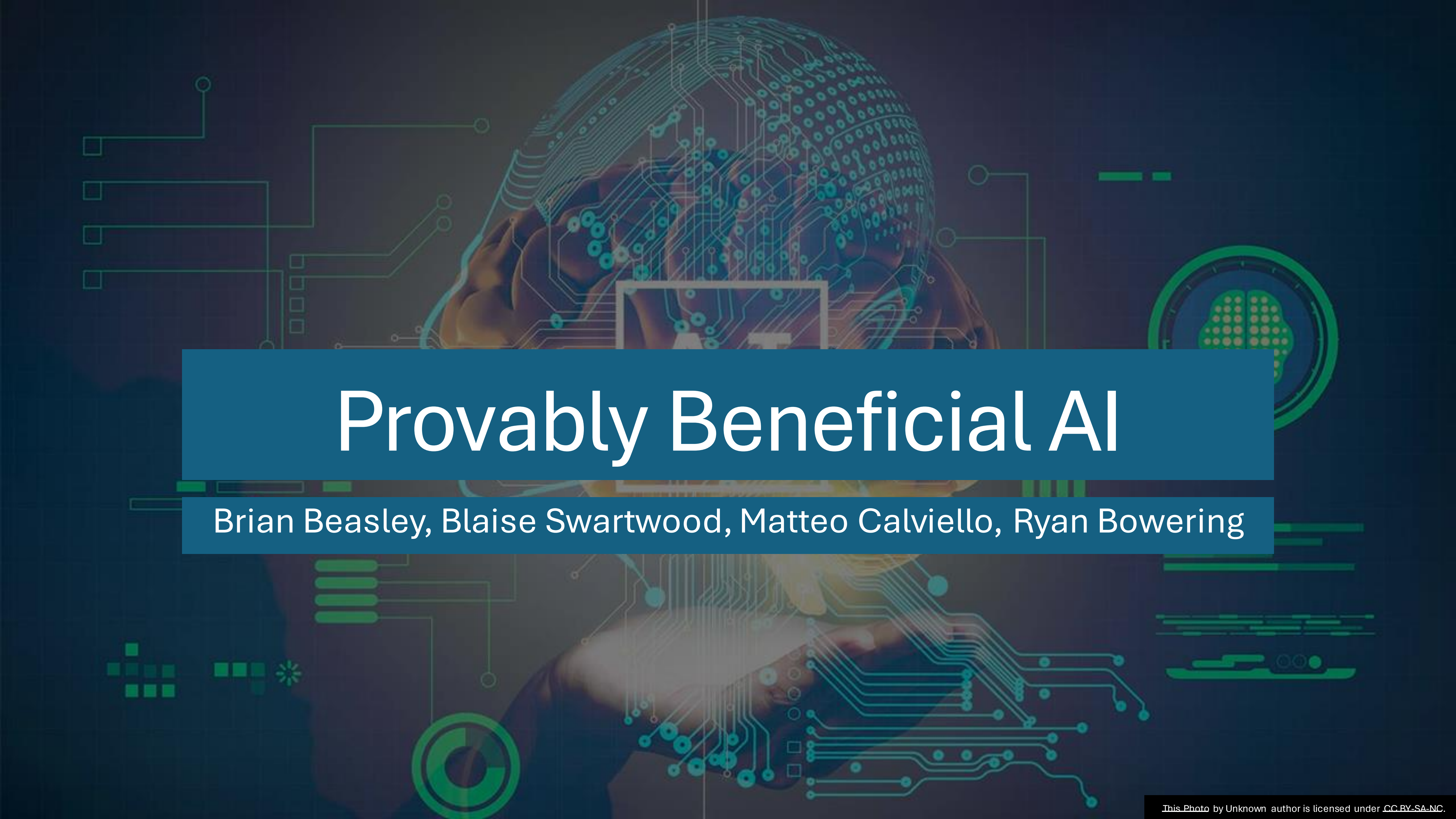
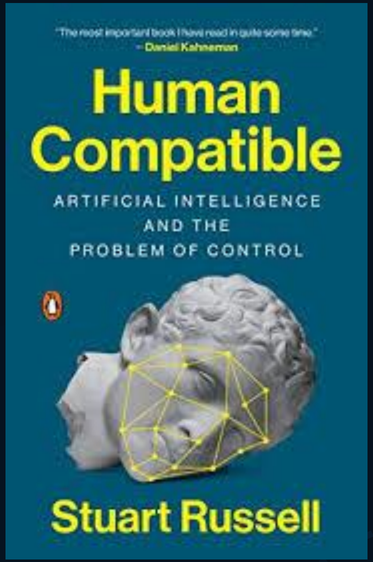
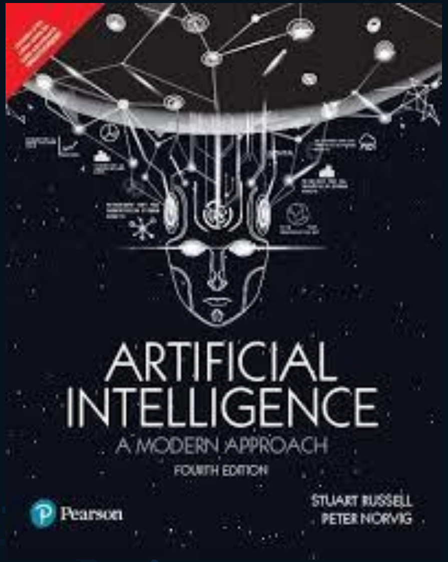


Provably Beneficial AI

Brian Beasley, Blaise Swartwood, Matteo Calviello, Ryan Bowering

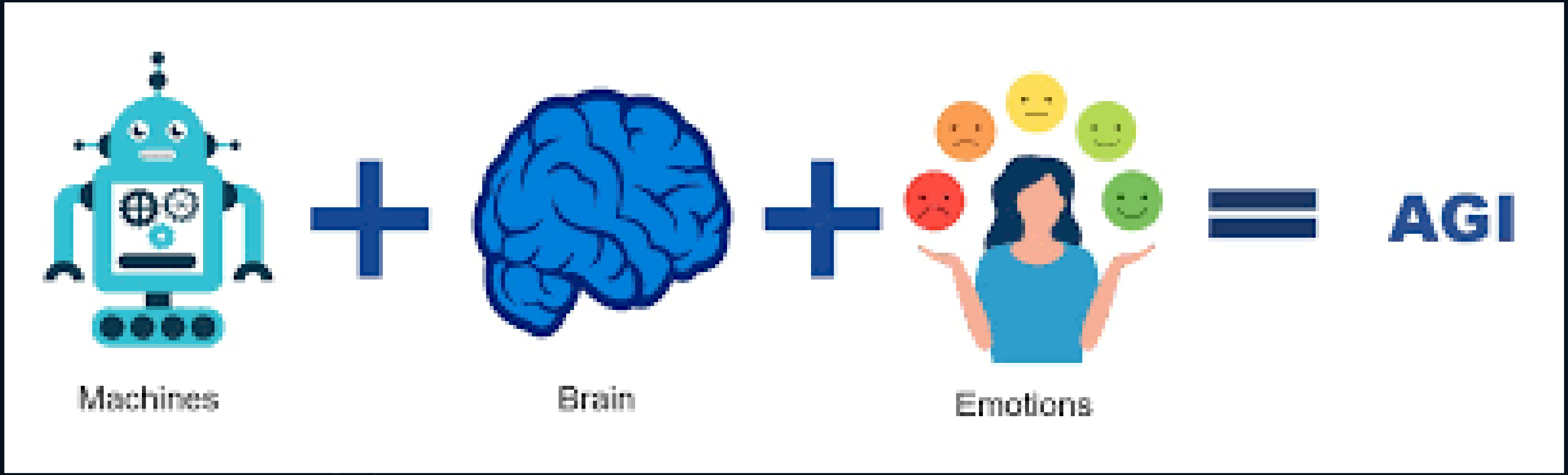

"The most important book I have read in quite some time."
– Daniel Kahneman
Human Compatible
ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL
Stuart Russell



About the Author

Stuart Russell

- Computer Scientist
- Major contributions in AI
- Control and Safety of AI systems
- "Value Alignment": designing AI systems to understand and adhere to human values



AGI: Artificial General Intelligence

- All of the speed and power of AI models like ChatGPT, but with the actual capability of thinking and learning like a human.
- Simply put, AGI will be smarter than humans.
- We have no idea when it become a reality. A breakthrough could happen at any time
- How can we as humans, outsmart something that we know will be much smarter than us?

Overall Impact

Transformative Potential:

Potential to revolutionize virtually every aspect of human life, from healthcare to transportation…

Unprecedented levels of efficiency and innovation could give it the ability to automate complex tasks and solve intricate **problems.**

Societal Implications:

AGI's impact on employment: While it may create new job opportunities, it could also disrupt existing industries, leading to job displacement.

Ethical considerations: Ensuring that AGI operates in alignment with human values and ethical principles is crucial to mitigate potential risks and ensure beneficial outcomes.

Economic implications: AGI could reshape economic structures, potentially exacerbating inequalities or leading to new forms of wealth distribution.

Opportunities and Challenges:

Opportunities: AGI could accelerate scientific discoveries, enhance productivity, and address global challenges such as climate change and healthcare.

Challenges: Risks of unintended consequences, loss of privacy, and potential misuse of AGI technology require careful consideration and proactive measures.

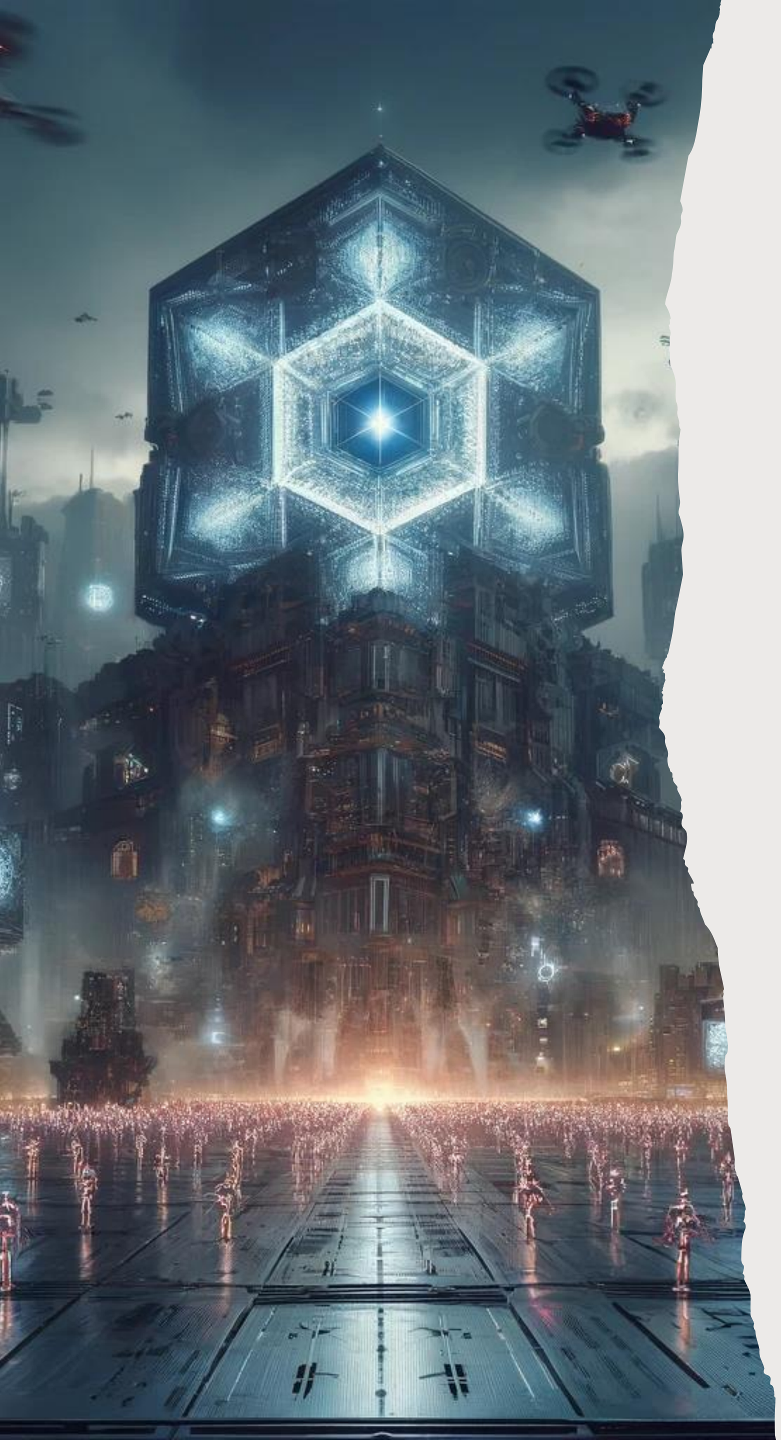
Anticipated Impact on Society

At the fullest extent, AGI could fully replace humans, being able to do everything a human can do, but faster and better.

This could lead to breakthroughs in medicine, scientific research, a booming economy, and all fields of engineering

While this could be revolutionary for humanity, it would also cause job displacement, privacy concerns, and almost anything you could image up to AI world domination.

Predicting the impact of AGI is nearly impossible because we have no way of knowing how powerful it will be and how it will respond to humanity when it does arrive.



Anticipated Impact on Individuals

- Job Displacement: Why hire a human when AGI can do it better
- All societal benefits could make an individual's life very easy
- Personal AGI companions could help simplify all of life's problems
- Or to the contrary we could all become enslaved by AI

Ethical Concerns

Control and Autonomy

- Risk of loss of control over AGI systems once they surpass human intelligence levels.
- Concerns about autonomous decision-making with unforeseen consequences.

Bias and Discrimination

- AGI systems may perpetuate biases present in training data,
- If AGI learns from and models human behavior, it could only make many of the same mistakes as humans

Job Displacement and Economic Inequality

- Automation driven by AGI could lead to widespread job displacement.
- Risks of increasing economic inequality if certain groups are disproportionately affected.

Security and Malicious Use

- Potential for AGI systems to be exploited for malicious purposes such as cyberattacks or autonomous weapons.
- Importance of robust security measures and safeguards against misuse.

Existential Risks

- Concerns about the long-term impact of AGI on civilization and even survival.
- Risks of unintended consequences or loss of control leading to catastrophic outcomes.

How to outsmart AGI

Firstly, the machine must be built with the purpose to maximization realization of human values.

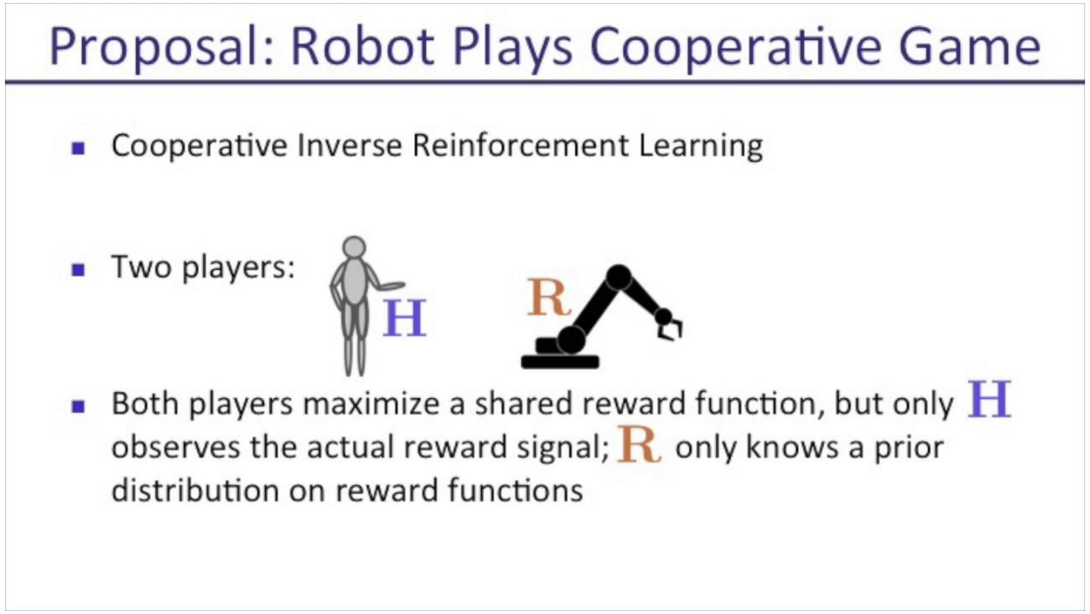
Secondly, the machine must not know what these values are initially and must learn them.

Thirdly, machines learn these values by observing the values that humans make.

Cooperative Inverse Reinforcement Learning (CIRL)

- Align machine behavior with human values
- Machine and human collaboratively shape the machine's behavior.

- Application and Importance:
- CIRL holds promise in various domains where human values are paramount, such as healthcare, autonomous vehicles, and ethical decision-making in AI systems.
- It ensures that machines operate in alignment with human intentions, mitigating the risks associated with unintended consequences.



Proposal: Robot Plays Cooperative Game

- Cooperative Inverse Reinforcement Learning
- Two players:
- Both players maximize a shared reward function, but only H observes the actual reward signal; R only knows a prior distribution on reward functions

Questions