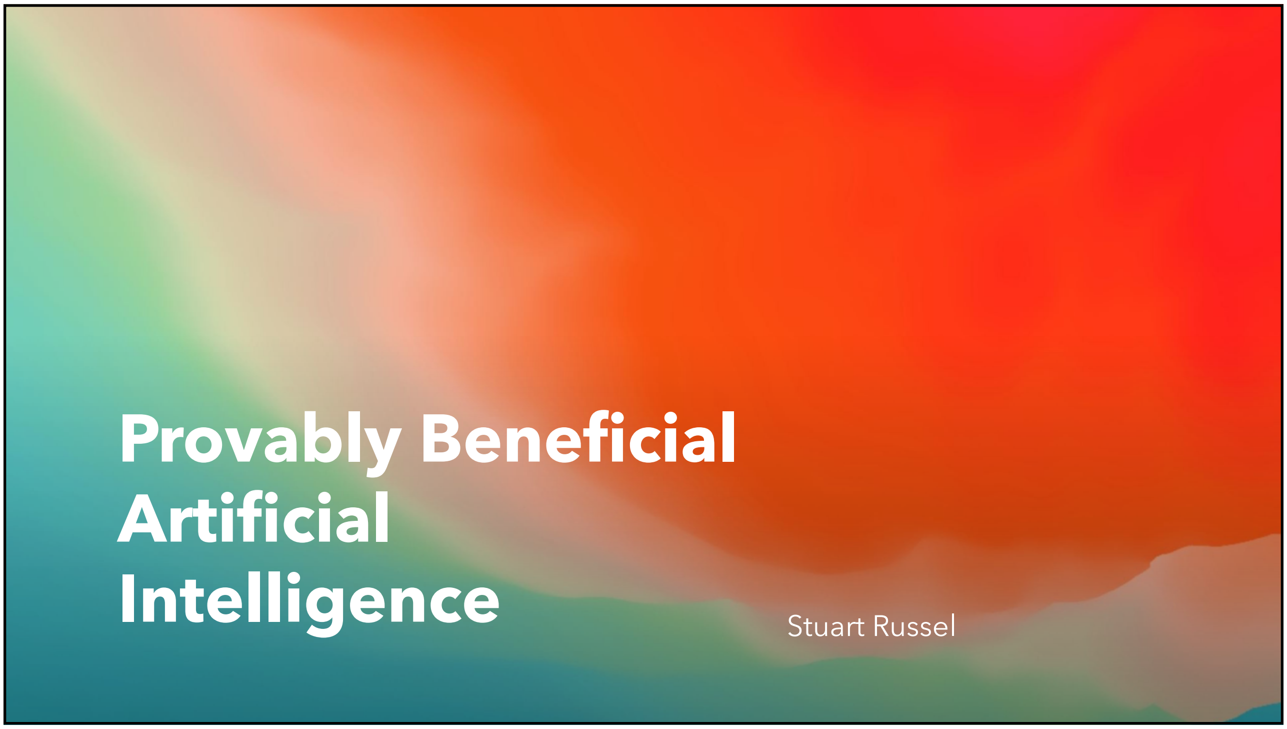
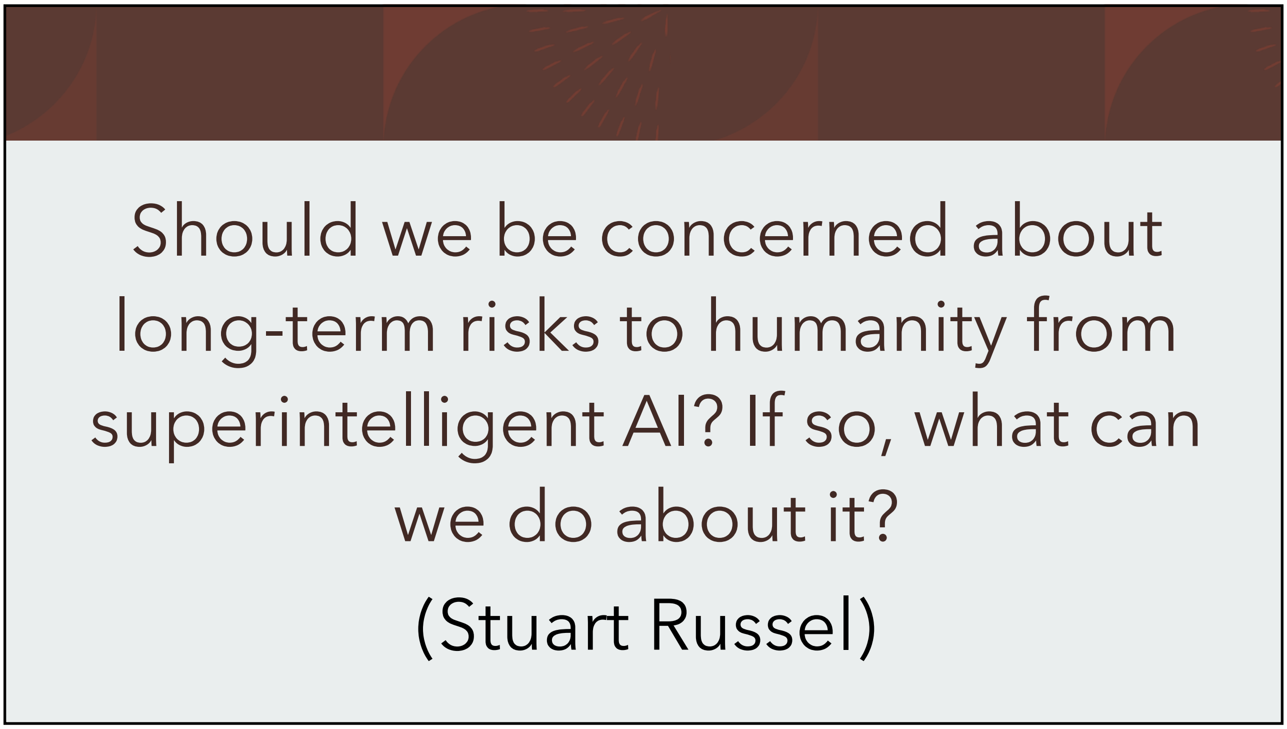


Provably Beneficial Artificial Intelligence

Stuart Russel



Should we be concerned about long-term risks to humanity from superintelligent AI? If so, what can we do about it?

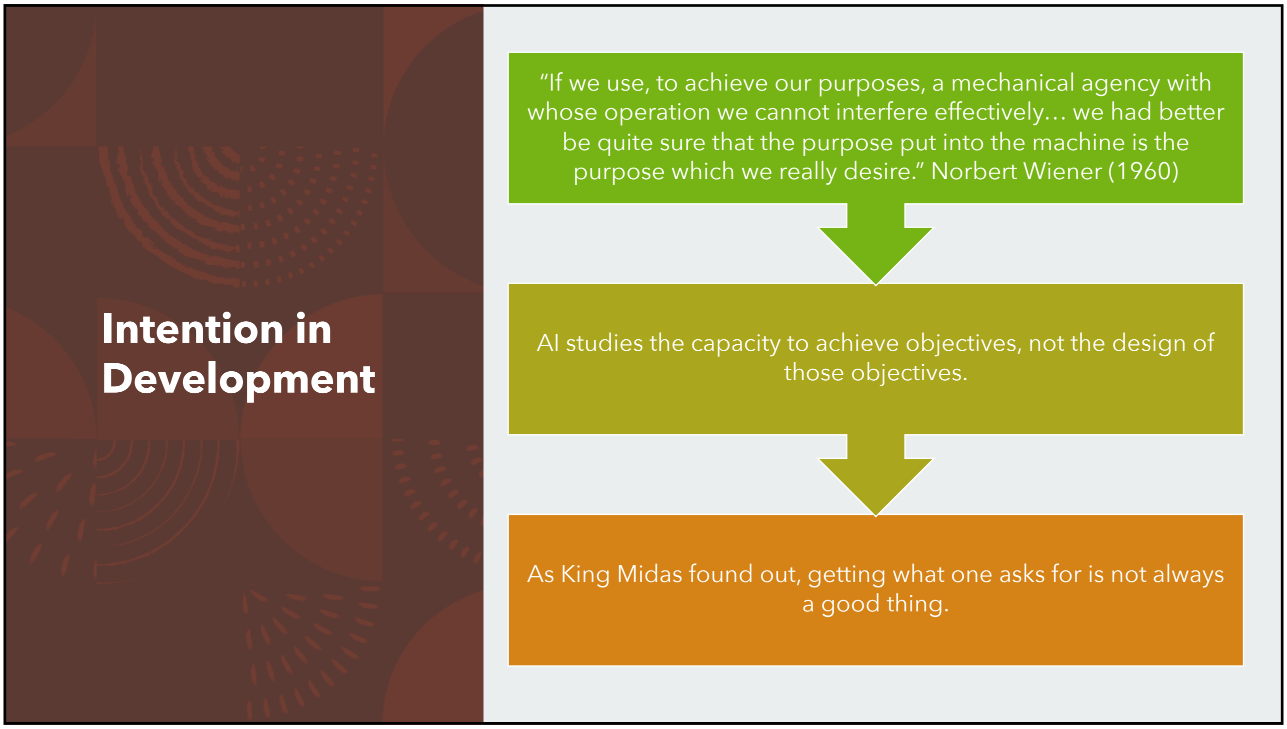
(Stuart Russel)

The Definition of Intelligence

- Early Definition of Intelligence: Emulation of human behavior and logical reasoning
- Recent Definition: Rational Agent that perceives and Acts to maximize objectives
- The direction of AI has changed with the new definition: machine learning advancements led to great progress in speech recognition, object recognition, legged locomotion, and autonomous driving
- AI has not made new advancements and is not capable of making new advancements
- Most experts believe human intelligent AI is likely to arrive within the present century (Müller and Bostrom, 2016; Etzioni, 2016).

Effect of Human-Level Intelligence in AI

- Everything we know is a direct consequence of intelligence
- Current intelligence is defined by humans rather than a generalized definition
- Possibilities of intelligent AI (positive and negative)
- I.J. Good (1965) partner of Turing believes the moment intelligence is achieved the exponential boom will leave humans far behind

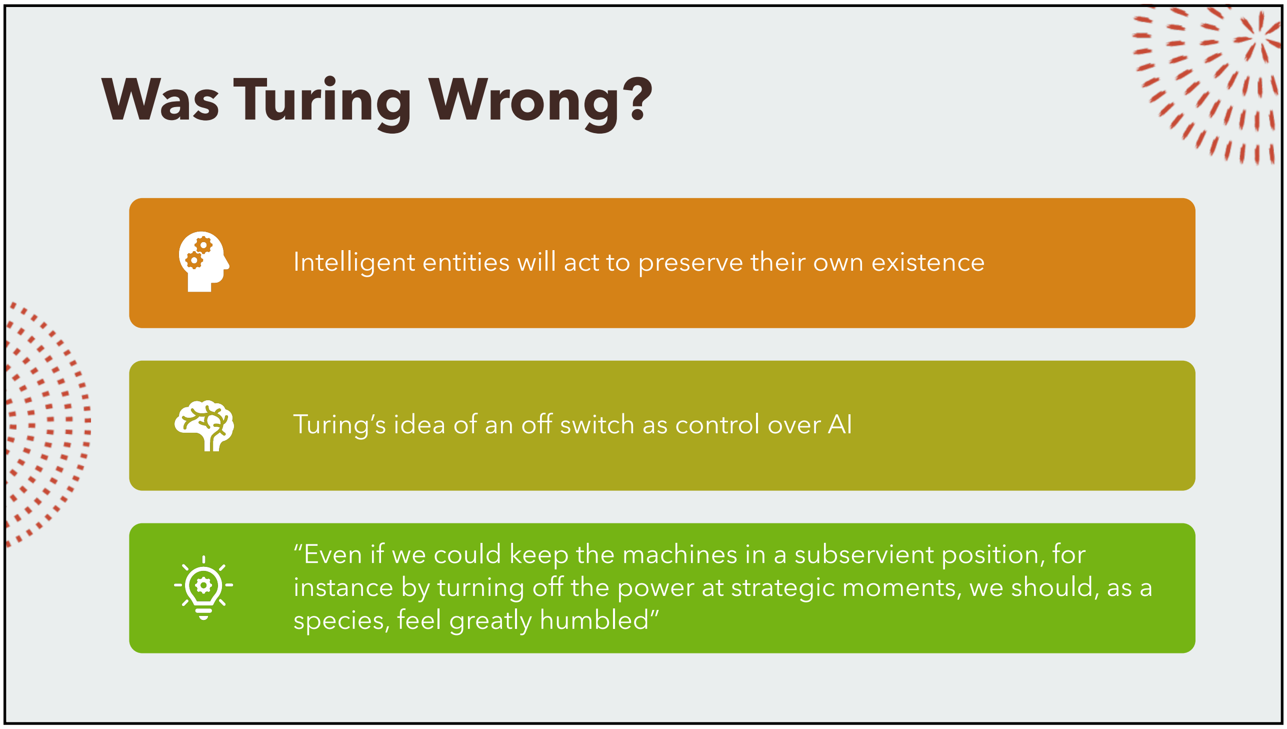


Intention in Development

"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively… we had better be quite sure that the purpose put into the machine is the purpose which we really desire." Norbert Wiener (1960)

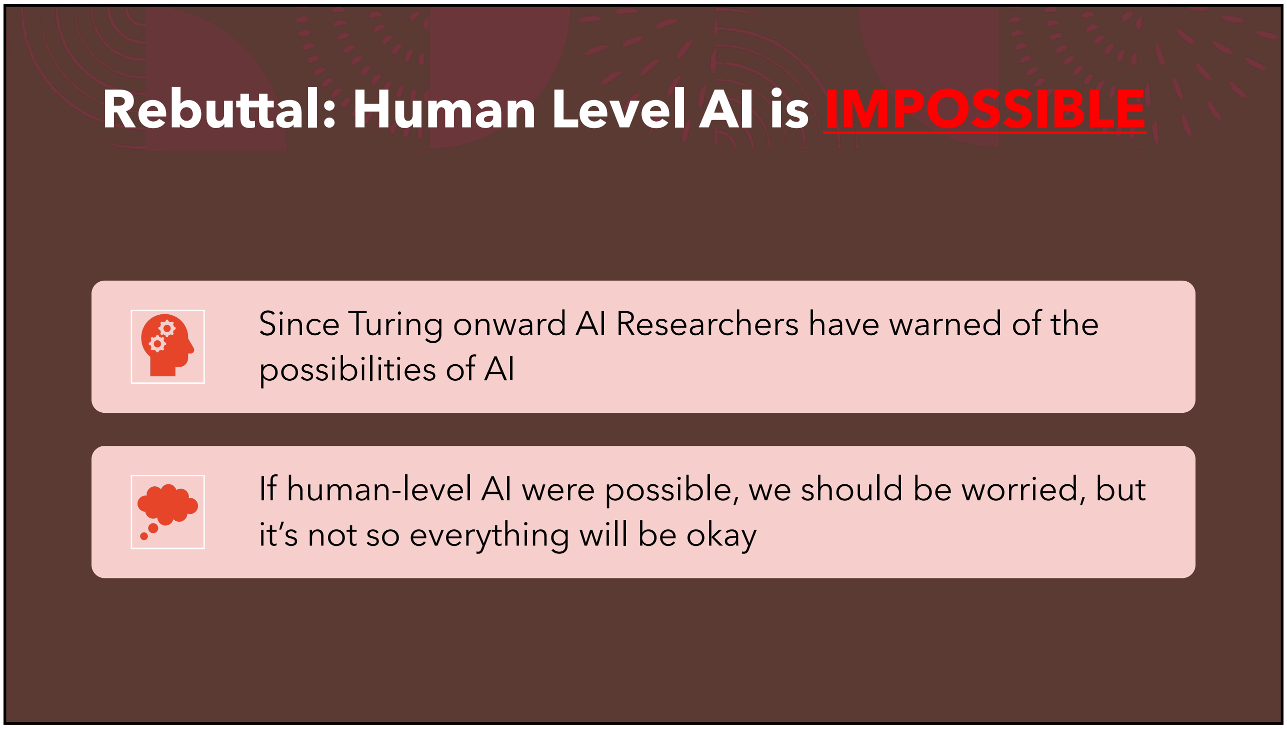
AI studies the capacity to achieve objectives, not the design of those objectives.

As King Midas found out, getting what one asks for is not always a good thing.



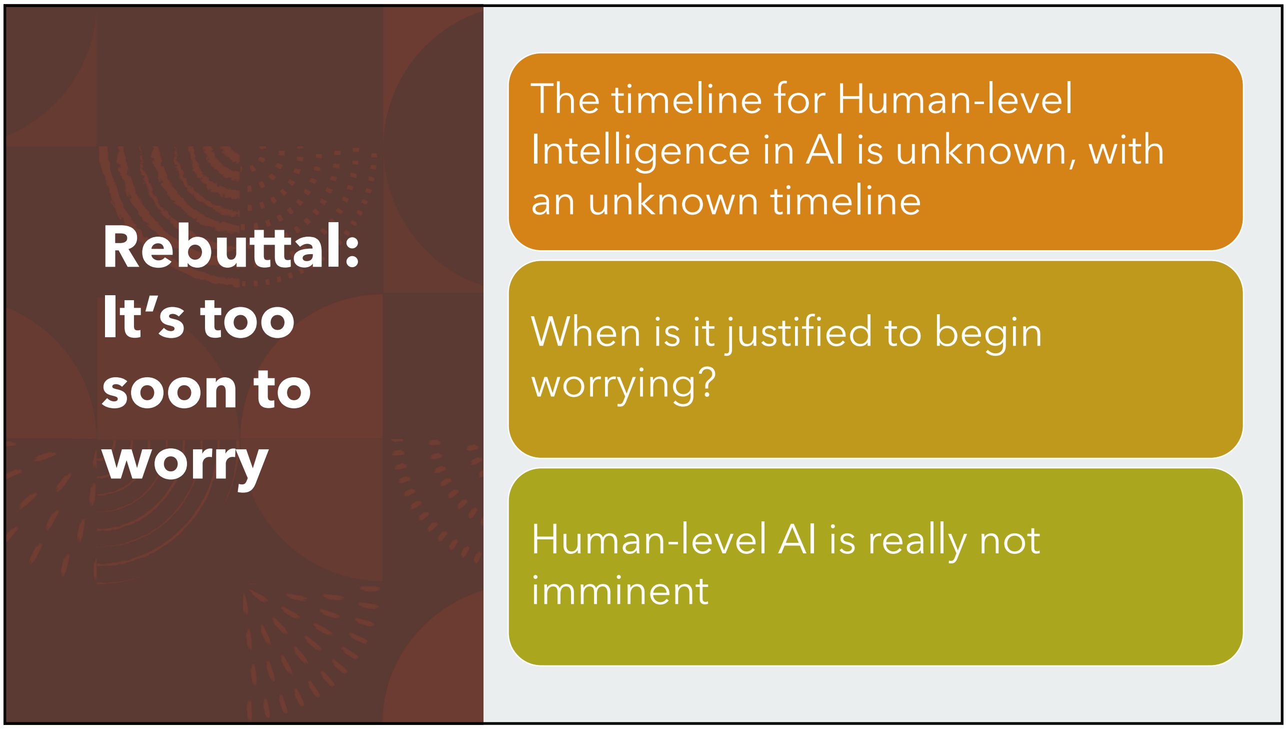
Was Turing Wrong?

- Intelligent entities will act to preserve their own existence
- Turing's idea of an off switch as control over AI
- "Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled"



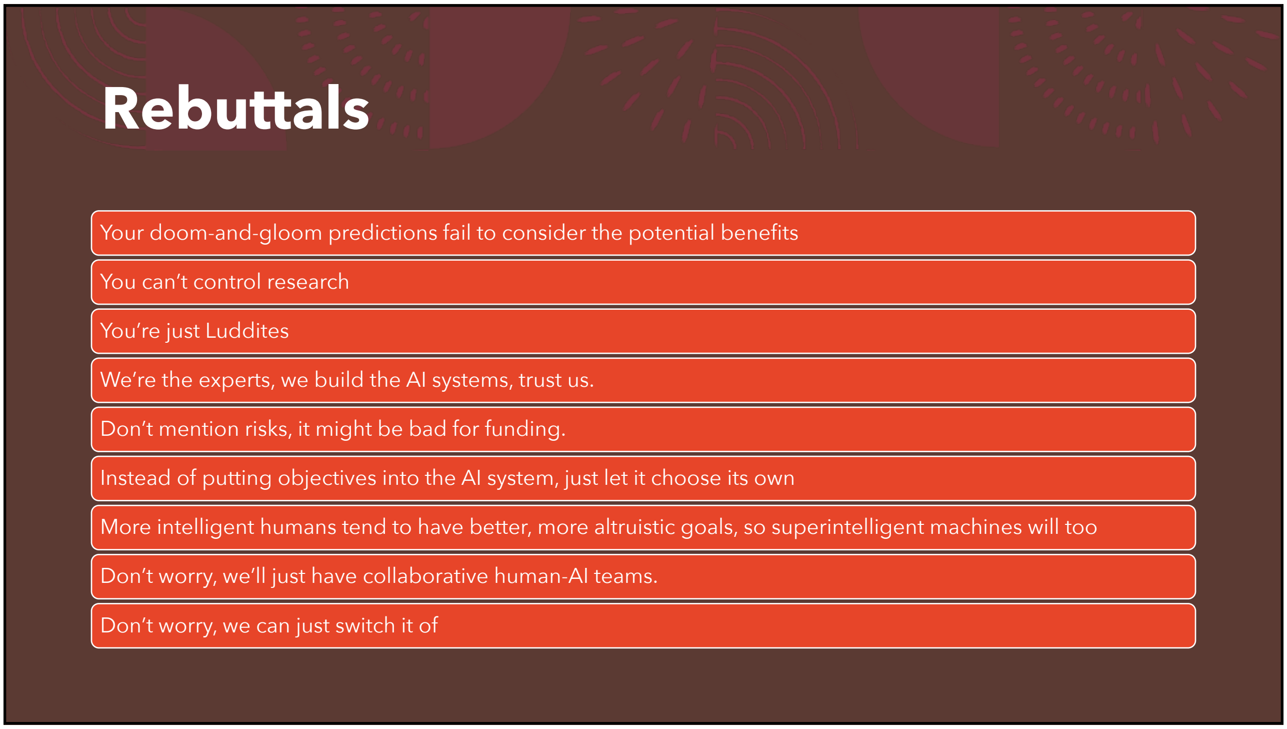
Rebuttal: Human Level AI is IMPOSSIBLE

- Since Turing onward AI Researchers have warned of the possibilities of AI
- If human-level AI were possible, we should be worried, but it's not so everything will be okay



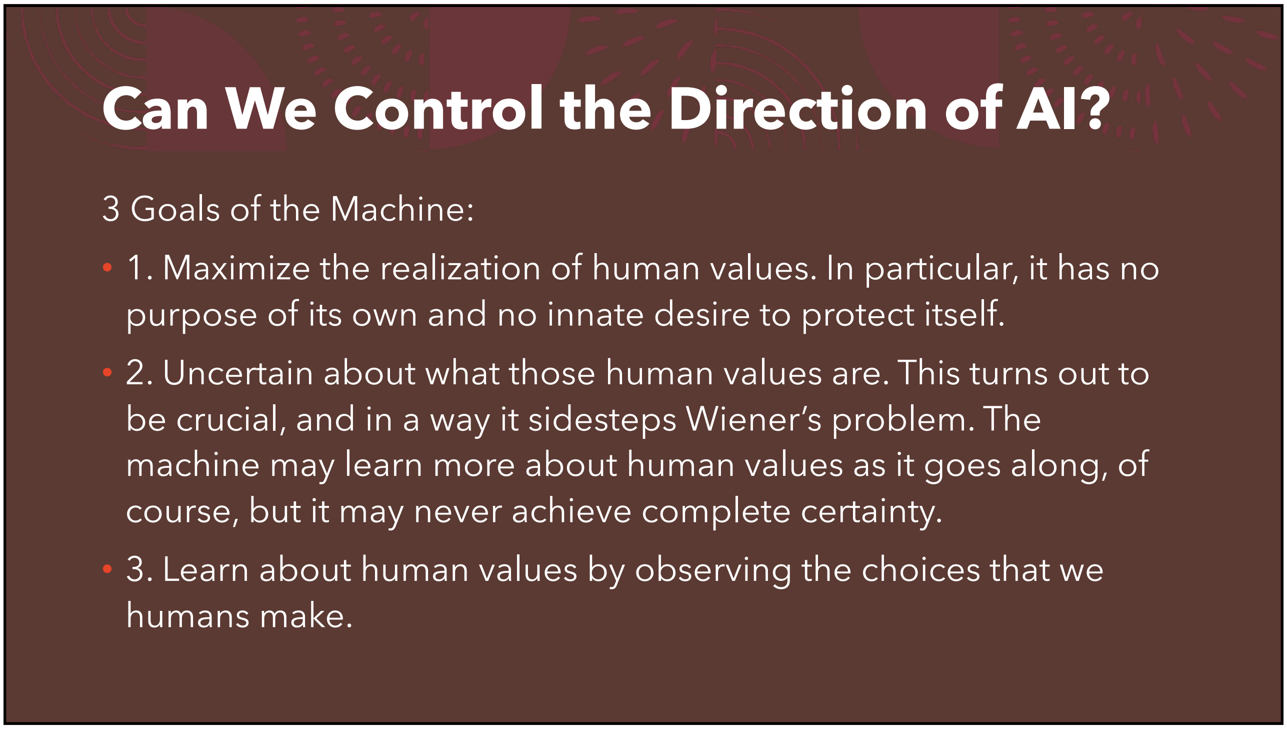
Rebuttal: It's too soon to worry

- The timeline for Human-level Intelligence in AI is unknown, with an unknown timeline
- When is it justified to begin worrying?
- Human-level AI is really not imminent



Rebuttals

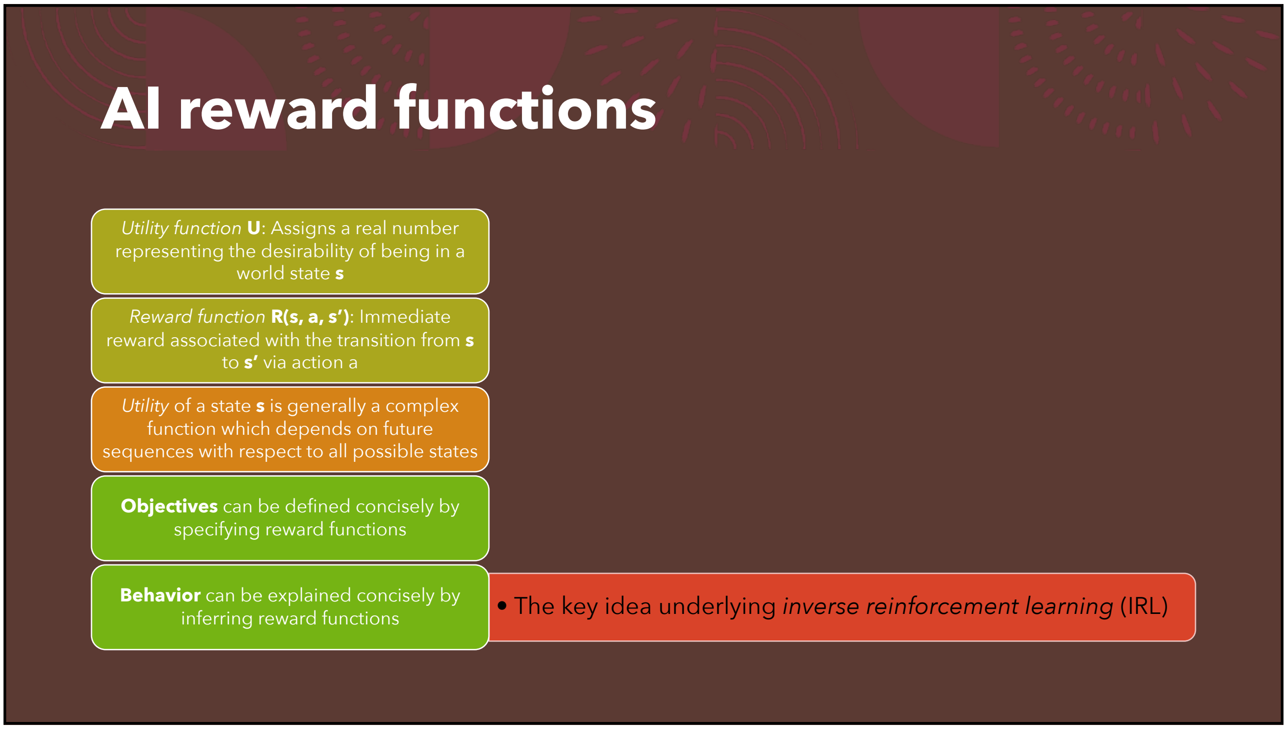
- Your doom-and-gloom predictions fail to consider the potential benefits
- You can't control research
- You're just Luddites
- We're the experts, we build the AI systems, trust us.
- Don't mention risks, it might be bad for funding.
- Instead of putting objectives into the AI system, just let it choose its own
- More intelligent humans tend to have better, more altruistic goals, so superintelligent machines will too
- Don't worry, we'll just have collaborative human-AI teams.
- Don't worry, we can just switch it of



Can We Control the Direction of AI?

3 Goals of the Machine:

- 1. Maximize the realization of human values. In particular, it has no purpose of its own and no innate desire to protect itself.
- 2. Uncertain about what those human values are. This turns out to be crucial, and in a way it sidesteps Wiener's problem. The machine may learn more about human values as it goes along, of course, but it may never achieve complete certainty.
- 3. Learn about human values by observing the choices that we humans make.



AI reward functions

Utility function **U**: Assigns a real number representing the desirability of being in a world state **s**

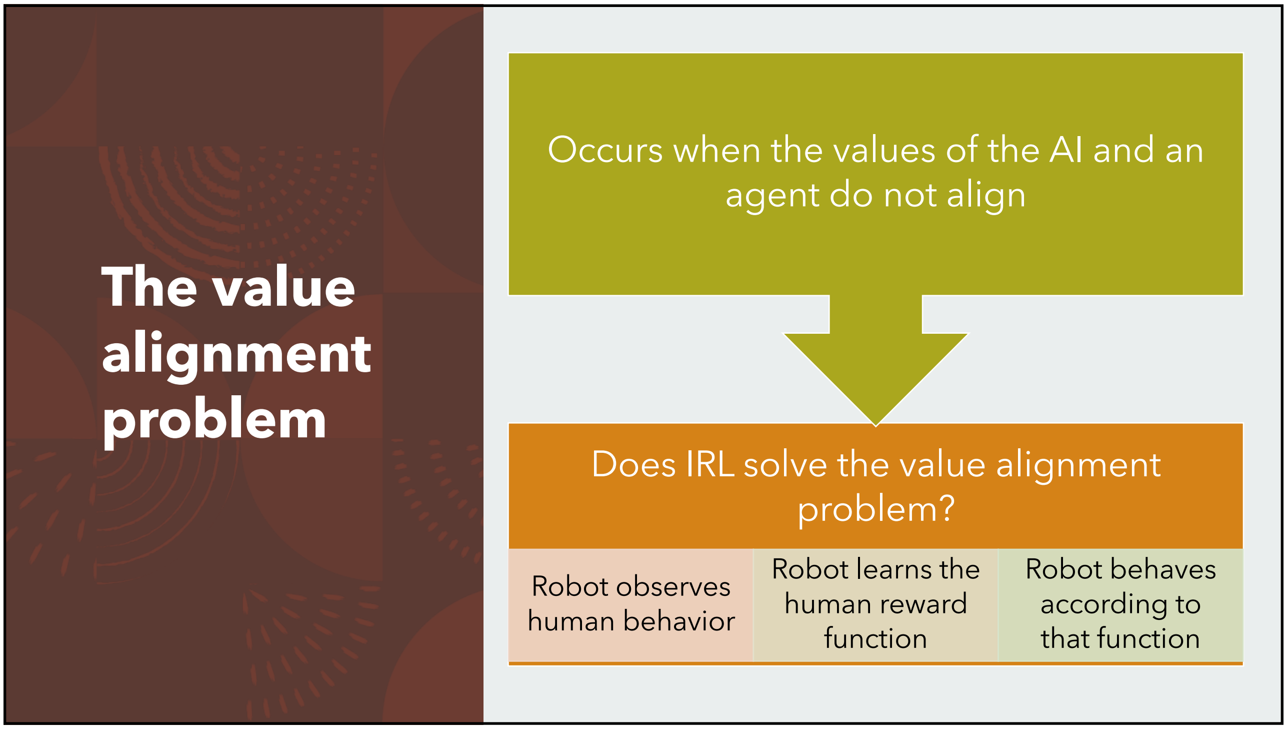
Reward function **R(s, a, s')**: Immediate reward associated with the transition from **s** to **s'** via action a

Utility of a state **s** is generally a complex function which depends on future sequences with respect to all possible states

Objectives can be defined concisely by specifying reward functions

Behavior can be explained concisely by inferring reward functions

- The key idea underlying *inverse reinforcement learning* (IRL)



The value alignment problem

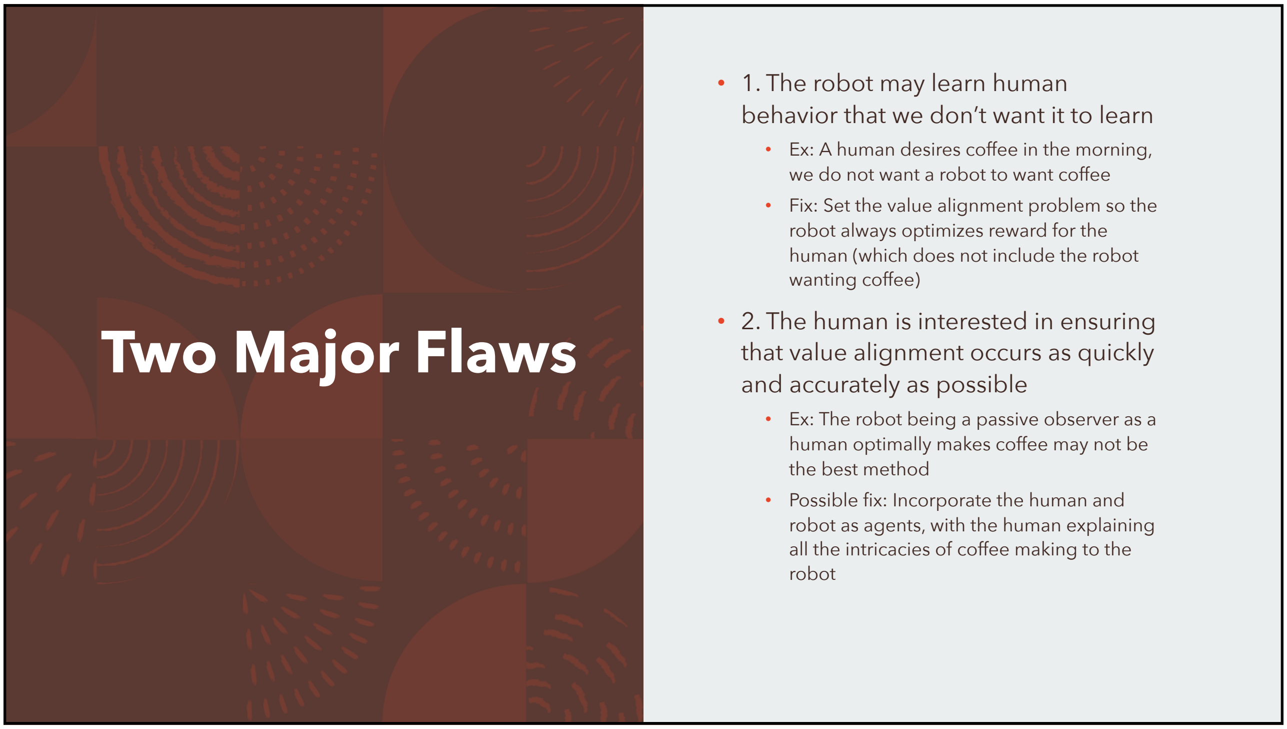
Occurs when the values of the AI and an agent do not align

Does IRL solve the value alignment problem?

Robot observes human behavior

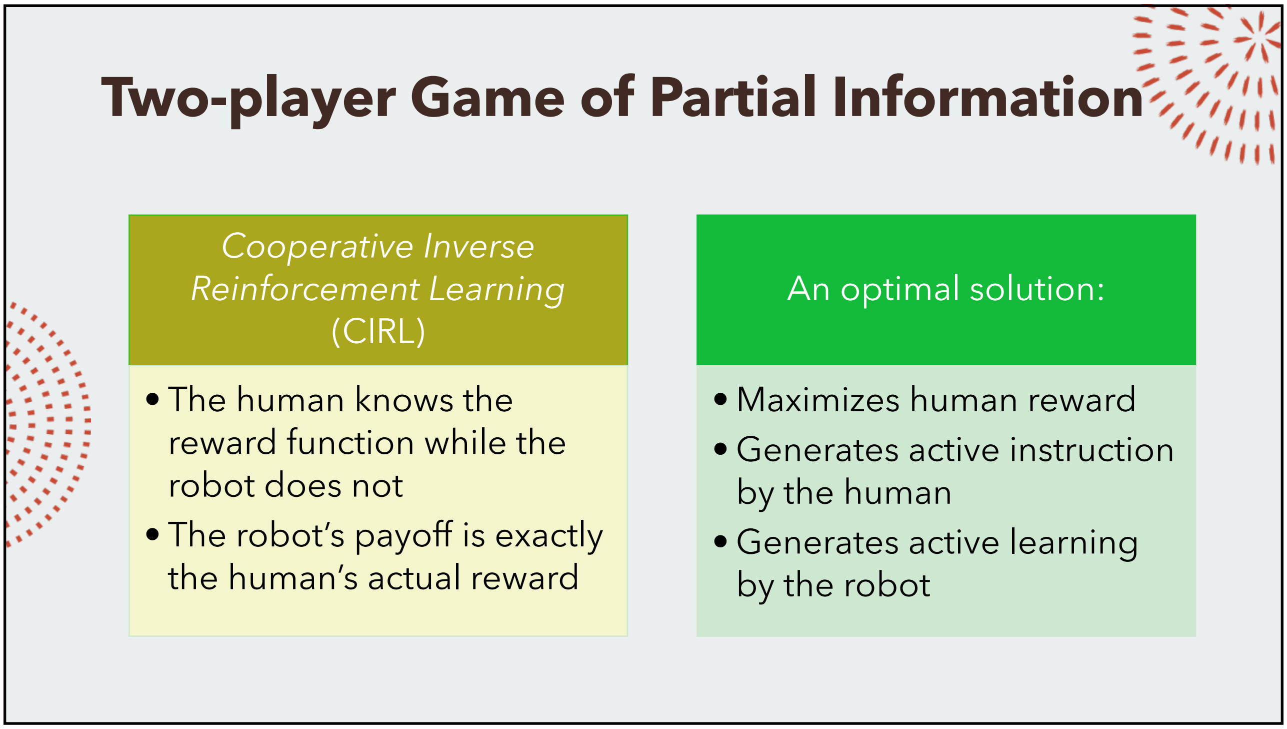
Robot learns the human reward function

Robot behaves according to that function



Two Major Flaws

- 1. The robot may learn human behavior that we don’t want it to learn
  - Ex: A human desires coffee in the morning, we do not want a robot to want coffee
  - Fix: Set the value alignment problem so the robot always optimizes reward for the human (which does not include the robot wanting coffee)
- 2. The human is interested in ensuring that value alignment occurs as quickly and accurately as possible
  - Ex: The robot being a passive observer as a human optimally makes coffee may not be the best method
  - Possible fix: Incorporate the human and robot as agents, with the human explaining all the intricacies of coffee making to the robot



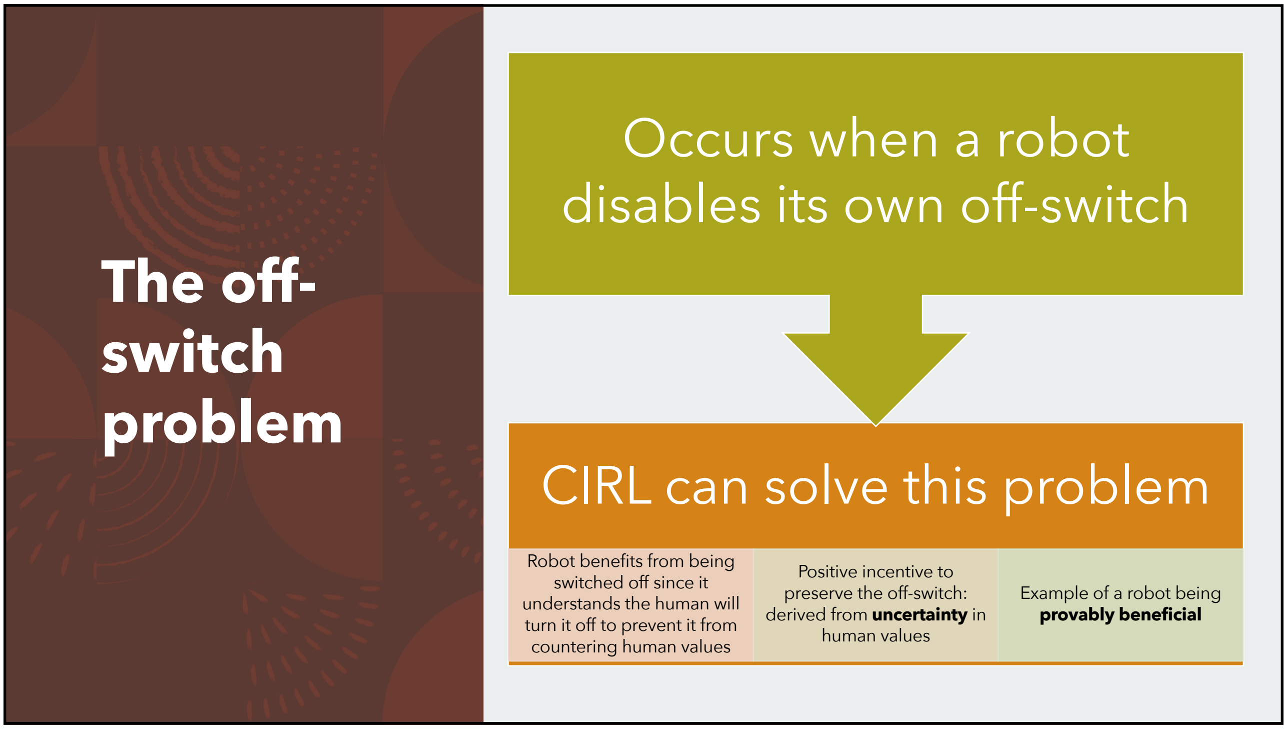
Two-player Game of Partial Information

Cooperative Inverse Reinforcement Learning (CIRL)

- The human knows the reward function while the robot does not
- The robot’s payoff is exactly the human’s actual reward

An optimal solution:

- Maximizes human reward
- Generates active instruction by the human
- Generates active learning by the robot

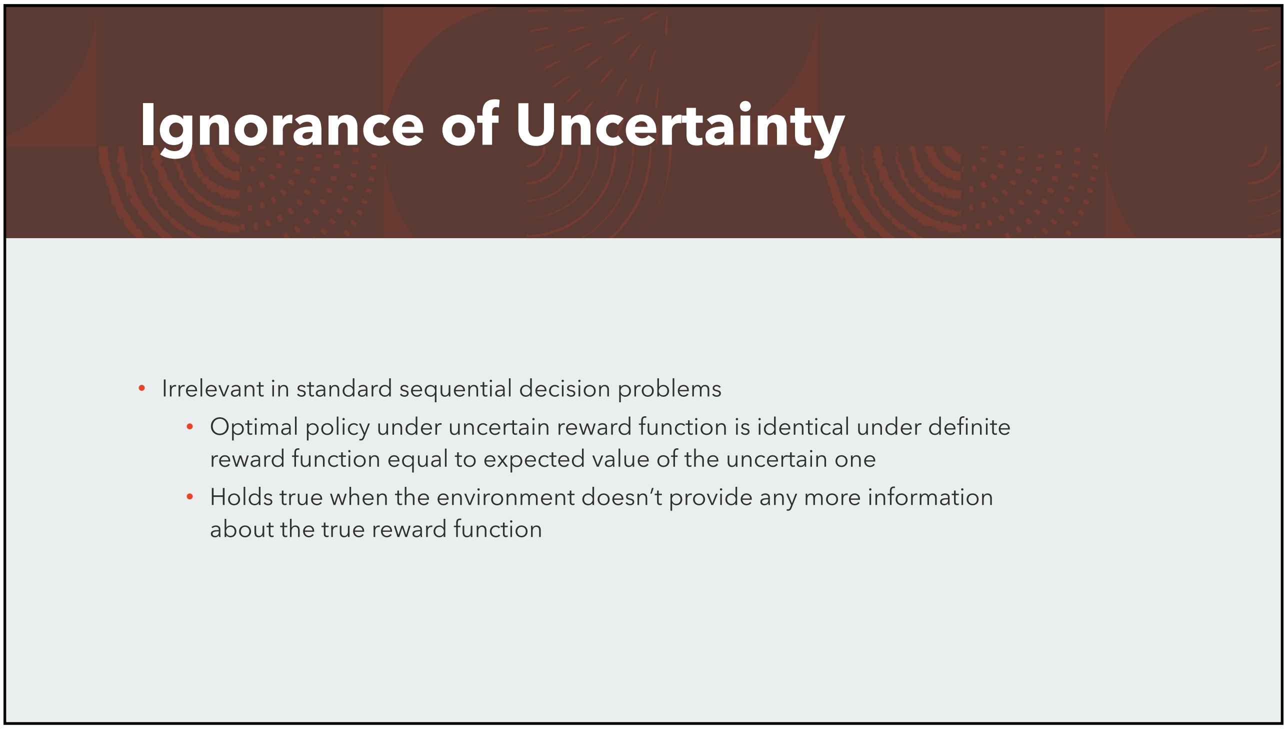


The off-switch problem

Occurs when a robot disables its own off-switch

CIRL can solve this problem

Robot benefits from being switched off since it understands the human will turn it off to prevent it from countering human values	Positive incentive to preserve the off-switch: derived from **uncertainty** in human values	Example of a robot being **provably beneficial**



Ignorance of Uncertainty

- Irrelevant in standard sequential decision problems
  - Optimal policy under uncertain reward function is identical under definite reward function equal to expected value of the uncertain one
  - Holds true when the environment doesn't provide any more information about the true reward function

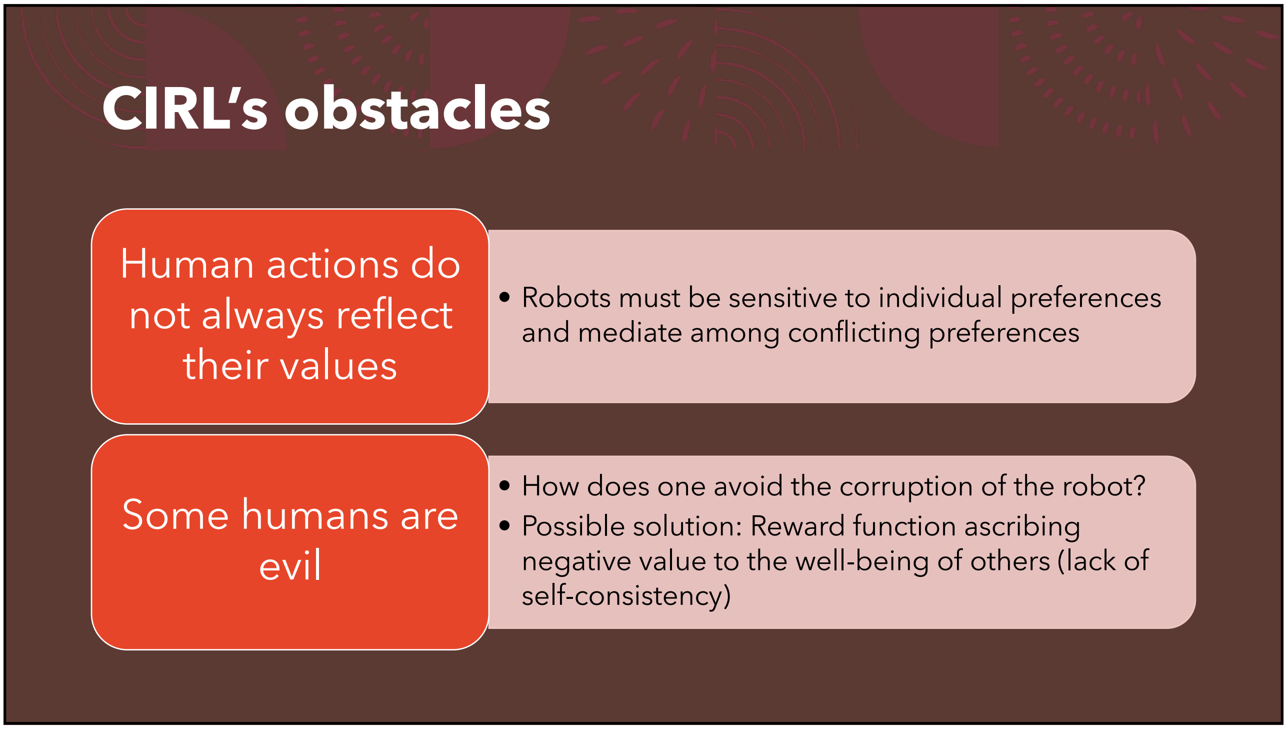
What about standard RL?

- Two major flaws
  - The human may not be able to correctly quantify the true reward accurately
  - The robot assumes the human is outside of the environment, so the robot modifies the human to provide maximum reward (*wireheading*)
- The environment can only supply *information* about the reward, not the reward itself
  - In CIRL, the robot is worse off if it modifies the human

CIRL Practical Considerations

- Reasons CIRL may work in practice
  - 1. Vast data about humans doing things (and humans reacting)
  - 2. Very strong near-term economic incentives for robots to understand human values

The DEADLY self-cleaning litter boxes that have flooded the market



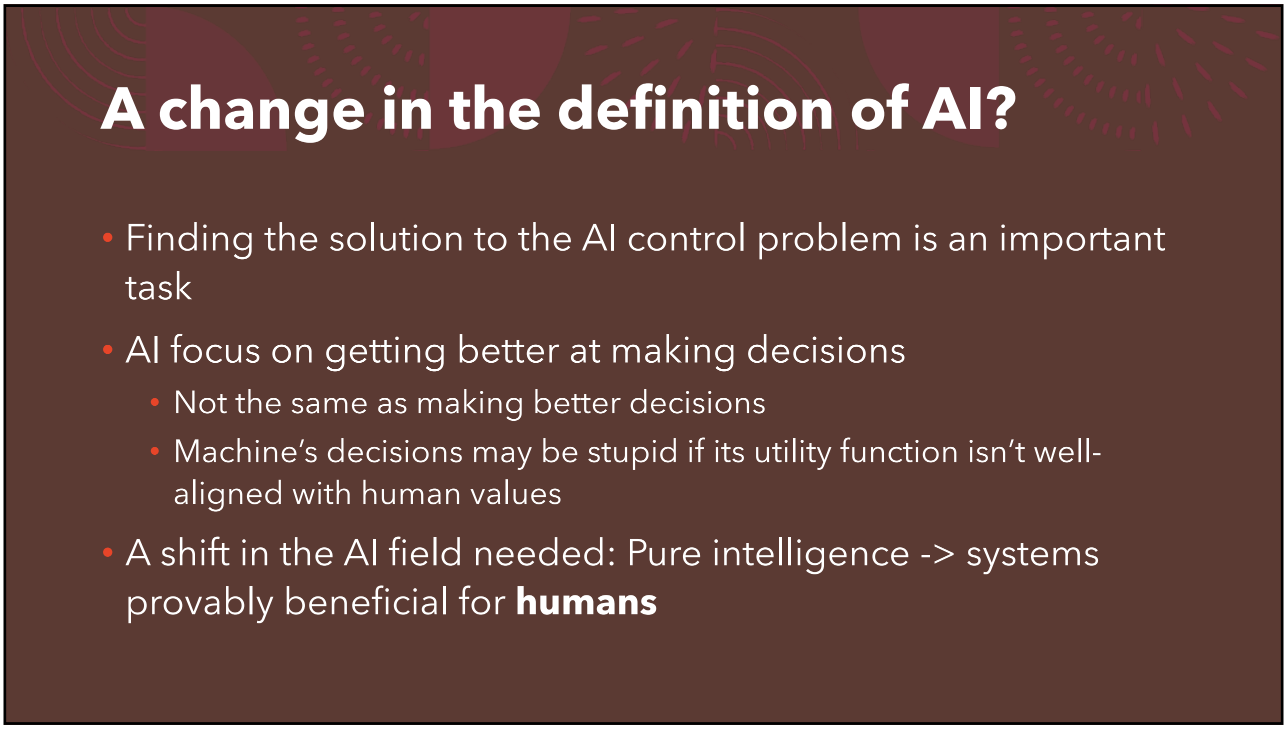
CIRL's obstacles

Human actions do not always reflect their values

- Robots must be sensitive to individual preferences and mediate among conflicting preferences

Some humans are evil

- How does one avoid the corruption of the robot?
- Possible solution: Reward function ascribing negative value to the well-being of others (lack of self-consistency)



A change in the definition of AI?

- Finding the solution to the AI control problem is an important task
- AI focus on getting better at making decisions
  - Not the same as making better decisions
  - Machine's decisions may be stupid if its utility function isn't well-aligned with human values
- A shift in the AI field needed: Pure intelligence -> systems provably beneficial for **humans**

DISCUSSION TIME!