




Provably Beneficial Artificial Intelligence

Alan Zhang, Lyra Lee, Michael Thede,
Wonhee Yu, Yulin Zhou



Introduction

Stuart Russell

- UC Berkeley CS Professor
- Notable contributions to AI research



Published in March 2022 for the 27th International Conference on Intelligent User Interfaces

Discusses AI Risks and his proposal to tackle them

AI Risks - AI Control

“If a machine can think, it might think more intelligently than we do ...” - Turing

- How to control something much more intelligent than humans are
- Need to identify a specific issue so research can be focused towards it
- Spontaneous evil is unrealistic in the real world

AI Risks - Value Alignment

- How can humans ensure AI is properly working towards humanity's best interests?
- Many related fields neglect this question and make assumptions
- Especially important if humans eventually cannot adjust AI learning

AI Risks - Off-Switch

- Intelligent beings will always take steps to ensure self-preservation
- An AI's purpose cannot be fulfilled if it is turned off

Rebuttals

Four Key Rebuttals to Concerns Regarding AI:

- 1) Dismissive Attitudes
- 2) Benefits Over Risk Emphasis
- 3) Policy and Research Control
- 4) Simplistic Solutions Critique

Solutions

Three core principles:

- 1) Maximize the realization of human values
- 2) Uncertain about what those human values are
- 3) Learn about human values by observing human choices

Inverse Reinforcement Learning (IRL)

Learns a reward function by observing the behavior of some other agent who is assumed to be acting in accordance with such a function

IRL does not solve the value alignment problem

1. We do not want robots to want coffee
2. We want coffee to be made quickly and accurately but this requires guidance from human

→ Achieving the value alignment does not fit in the standard IRL framework

Cooperative Inverse Reinforcement Learning (CIRL)

Two player game of partial information, human knows the reward function while the robot is not. Robot's payoff is exactly the human's actual reward

- Solves value alignment problem
- Checks all three core principles
- Solves off-switch problem

→ Some cases robot is “provably beneficial”

Practical Considerations

Workable reasons:

Vast amount of information about humans

Very strong, near-term economic incentives for robots

Difficulties:

Humans are irrational, inconsistent, diverse, evil, ...

Evil behavior?

Impact on Society / Individuals

Society:

- Paradigm shift on usage of AI
- Improper instruction now can result in less efficient AI down the road

Individuals:

- AI could advise less individual-focused solutions and hurt the user
- Too much optimization could lead to a breakdown of action

Evaluation/Thoughts

- Risks are realistic
- Solutions seem accurate in the idea of matching incentives
- Variation of preferences will be hard to train
 - Time intensive/cost incentive
- Ethical concerns for bad usage
 - No different from other tech

Discussion

- Additional thoughts or rebuttals to these risks?
- When will these risk become real, if ever?
- How would you ensure that AI is trained fairly to maximize return to public?
- Other ethical concerns that might arise with 'provably beneficial AI'?

Resources

<https://people.eecs.berkeley.edu/~russell/papers/russell-bbvabook17-pbai.pdf>

(<https://dl.acm.org/doi/10.1145/3490099.3519388>)

<https://time.com/collection/time100-ai/6309044/stuart-russell/>

<http://aima.cs.berkeley.edu/adoptions.html>