A Brief Summary of Research on Provably Beneficial AI

Stuart Russell
University of California, Berkeley
[joint work with Dylan Hadfield-Menell, Smitha Milli, Anca Dragan, Pieter Abbeel, Tom Griffiths]
Good AI systems

- Restricted systems (tool AI)
- Constraints on a smart system
- Value alignment
- other
Value alignment

- Inverse reinforcement learning (IRL)
- Cooperative IRL (CIRL): a two-player game with "human" and "robot"
  - Human "knows" the value function (usually acts according to it)
  - Robot doesn’t know it, but wants to maximize it
- Optimal solutions have these properties:
  - Robot has an incentive to ask questions first
  - Human has an incentive to teach the robot
    - Human behavior is “suboptimal”
    - So an IRL algorithm shouldn’t expect “optimal”
“If a machine can think, it might think more intelligently than we do, and then where should we be? 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. ...

[T]his new danger ... is certainly something which can give us anxiety.”

Alan Turing, 1951
A robot, given an objective, has an incentive to disable its own off-switch

“You can’t fetch the coffee if you’re dead”

How can we prevent this?

Answer: robot isn’t given an objective!

It must be uncertain about the true objective

The human will only switch off the robot if that leads to better outcomes for the true human objective

Theorem: it’s in the robot’s interest to allow it

cf non-negative value of information
Desired Behavior

Incorrignble Behavior

Non-Functional Behavior
Off-switch model

\[ U = U_a \]

\( a \)

\( w(a) \)

\( U = 0 \)

\( s \)

\( \neg s \)

\( U = 0 \)

\( U = U_a \)

\( a \)

\( a \text{ or } s \) preferred to \( w(a) \)

\( w(a) \) preferred to \( a \) or \( s \)
Increasing uncertainty

Increasing $H_{\text{suboptimality}}$

Increasing $R_{\text{uncertainty}}$

$R_{\text{incorrigible}}$

$R_{\text{corrigible}}$
Uncertainty in objectives

- *Irrelevant* in standard decision problems...
- *...Unless* the environment provides further information about objectives
  - E.g., observable human actions
  - A “reward signal” is a human action that provides *information not reward*
    - Avoids the wireheading problem
Value alignment contd.

- Humans are nasty, irrational, inconsistent, weak-willed, computationally limited, and heterogeneous
Center for Human-Compatible AI

... to reorient the general thrust of AI research towards provably beneficial systems
Current topics

- What is an instruction?
- What is an advising machine?
- Can we make safe question-answering systems of arbitrary ability?
- Extensions of CIRL to multiple humans and robots (possibly w/ global sharing)
- Safety margins when the robot may be unaware of some dimensions of U