Provably Beneficial Artificial Intelligence

Stuart Russell
University of California, Berkeley
Eventually, AI systems will make better* decisions than humans

- Taking into account more information, looking further into the future
Robots Could Replace Half Of All Jobs In 20 Years

By Timothy Torres, Tech Times | March 24, 6:56 PM

If we’re to believe University of Oxford associate professor Michael Osborne, then robots will replace 47 percent of all jobs by the year 2035.

If you want to stay employed by then, you better think about a career shift into software development, higher level management or the information sector. Those professions are only at a 10 percent risk of replacement by robots, according to Osborne. But
Most people will “work” improving each others’ lives
To add value and derive income, such work must be effective
We need to completely retool our education system and science base
WELCOME TO

UTOPIA

ENJOY YOUR JOURNEY
Artificial Intelligence could spell the end of the human race
We had better be quite sure that the purpose put into the machine is the purpose which we really desire

Norbert Wiener, 1960

King Midas, c540 BCE
Changing AI

- Standard AI (and many other fields):
  - Design systems that optimize a given objective
- Provably beneficial AI:
  - Design systems that behave in such a way that humans are happy with the results
  - Proposal: AI systems solve cooperative inverse reinforcement learning games
Basic principles

1. The robot’s only objective is to maximize the realization of individual human preferences
2. The robot is initially uncertain about what those preferences are
3. Human behavior provides information about human preferences
The off-switch problem

I must fetch the coffee

I can’t fetch the coffee if I’m dead

Therefore I must disable my off-switch

And Taser all other Starbucks customers

Image courtesy of Clearpath Robotics
… with uncertain objectives

The human might switch me off

But only if I’m doing something wrong

I don’t know what “wrong” is but I know I don’t want to do it

Therefore I should let the human switch me off

Image courtesy of Clearpath Robotics
... with uncertain objectives

Theorem: Such a robot is provably beneficial
Difficulties

Us

• Computationally limited
• Inconsistent preferences
• Internal conflict
• Nasty
Reasons for optimism

- Huge volume of data on human choices
Reasons for optimism

- Huge volume of data on human choices
- Strong economic incentives to get it right
Your wife called to remind you about dinner tonight

For your 20th anniversary, at 7pm

I did warn you, but you overrode my recommendation…

Don’t worry, I arranged for his plane to be delayed – some kind of computer malfunction.

He sends his profound apologies and is happy to meet you for lunch tomorrow

Wait! What? What dinner?

I can’t, I’m meeting the Secretary General at 7.30! How did this happen??

OK, but what am I going to do now? I can’t just tell him I’m too busy!!

Really? You can do that?!?
Welcome home! Long day?

Yes, terrible, not even time for lunch.

So you must be quite hungry!

Starving! Can you make me some dinner?

There’s something I need to tell you

There are humans in South Sudan in more urgent need of help. I am leaving now. Please make your own dinner.
Summary

- Rapid progress in AI is impacting society
- Prepare for major economic disruption
- Develop the theory and practice of provably beneficial AI