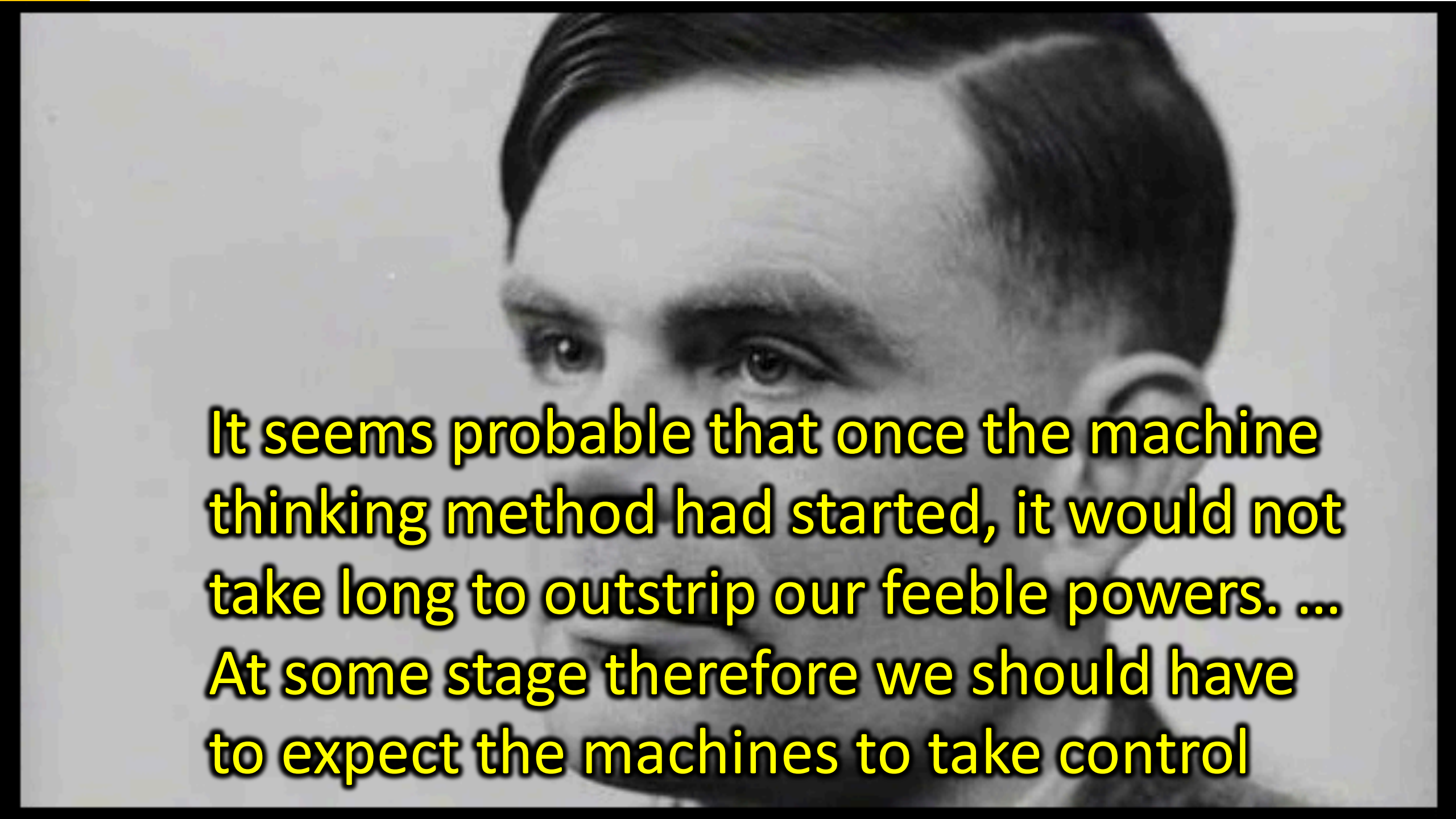


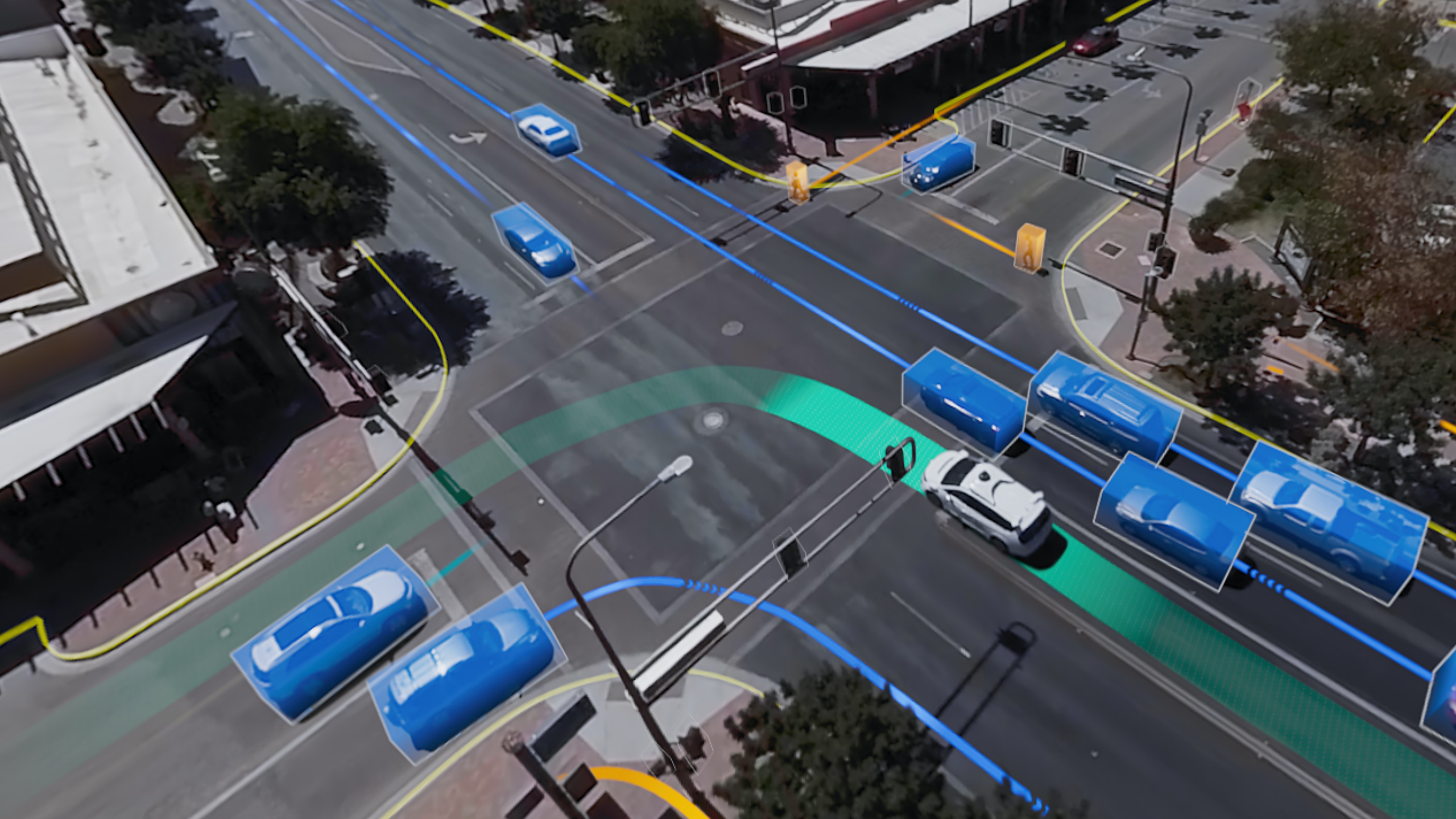


# How Not to Destroy the World with Artificial Intelligence

**Stuart Russell**  
**UC Berkeley**



**It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control**





# TaaS: Getting to Australia

**In 1800:** \$1,000,000,000, 10 years

- probably dead

**In 2020:** \$1,000, 1 day

- almost certainly alive

# XaaS

- Apply the same cost reduction to everything
  - Organizing and running a large conference
  - Building houses, schools, hospitals, roads
  - Teaching children, training surgeons
- Regional or global AI systems with a variety of physical extensions (legged/winged/wheeled)

# Benefits

- **Lift the living standards of everyone on Earth to a respectable level**
  - => 10x increase in world GDP
  - => \$13.5Q Net Present Value
- **Conflict to gain a bigger share of wealth will be like fighting over who has more digital copies of the newspaper**



# Eventually...

**AI systems will make better real-world decisions than humans**

**Turing's point: how do we retain power over entities more powerful than us, for ever?**





“Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.

**Stephen Hawking**



# Standard model of AI

(and control theory, statistics, operations research, economics)

**Machines that optimize an exogenously specified objective**

**But we cannot specify objectives completely and correctly**



**Third wish = please undo first two wishes**

# Social media catastrophe

Objective: maximize clickthrough

~~— learning what people want~~

= modifying people to be more predictable

The better the AI, the worse the outcome!

# How we got into this mess

- **Humans** are intelligent to the extent that **our** actions can be expected to achieve **our** objectives
- ~~**Machines** are intelligent to the extent that **their** actions can be expected to achieve **their** objectives~~
- **Machines** are *beneficial* to the extent that *their* actions can be expected to achieve *our* objectives

# New model: Provably Beneficial AI

1. Robot goal: satisfy human preferences\*
2. Robot is uncertain about human preferences
3. Human behavior provides evidence of preferences

The robot solves a formally defined assistance game

Optimal solutions:

defer to human, ask permission, allow self to be switched off

**The better the AI, the better the outcome!**

# 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**

# ... 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**

# ... with uncertain objectives



Θη 'υμαν =μειτ Σωιτχ μι οφ

Π<sub>1</sub> μπυτ = ωνλη ιφ ειμ +  
δοιγγ Συμθιγγ ρογγ

Π<sub>1</sub> ιδωντ νω ωατ ρογγ ιζ μπυτ  
αι δωντ ωαντ τυ δυ ιτ

ΣΠ Θηρφωρ Ι λετ θη +  
'υμαν σωιτχ μη οφ

**Theorem: Robot is provably beneficial**



# Extending the basic theory

## Many humans

=> connections to moral philosophy, economics

## Non-rational humans

=> connections to cognitive psychology, neuroscience

## Foundations

=> rebuild each area of AI (search, planning, RL, etc.)

## Applications

=> self-driving cars, digital assistants, personal robots

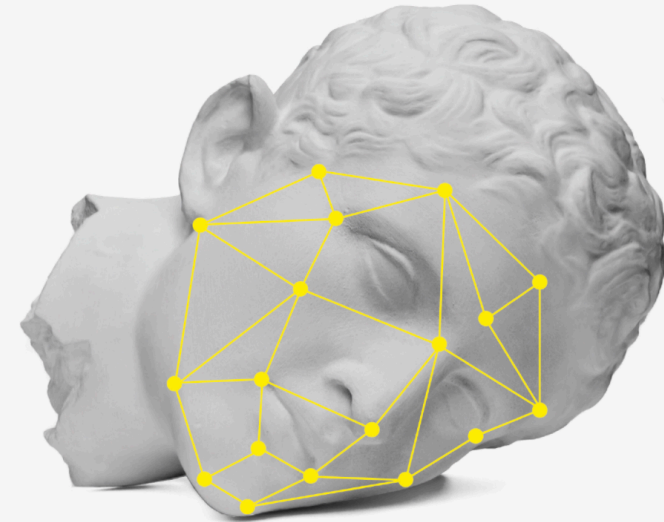
Stuart J. Russell  
HUMAN  
COMPATIBLE



AI and the Problem of Control

allen lane

STUART RUSSELL  
HUMAN  
COMPATIBLE



**Künstliche Intelligenz**

und wie der Mensch die Kontrolle über  
superintelligente Maschinen behält



# Summary

The standard model for AI leads to loss of human control over increasingly intelligent AI systems

Provably beneficial AI is possible and desirable  
It's not AI Ethics, it's AI

Problems of misuse and overuse are completely unsolved

