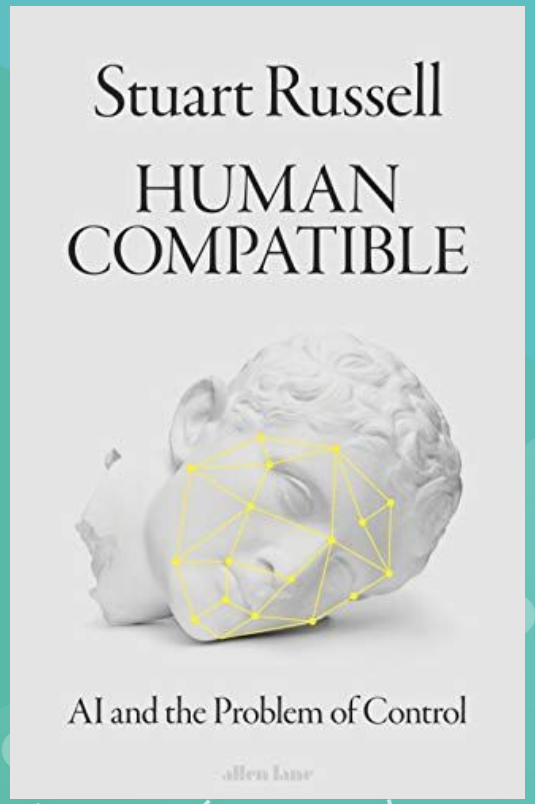
HUMAN COMPATIBLE

AI and the Problem of Control

A critical discussion by
Luciano Cavalcante Siebert



Preface

- Stuart Russel
  - Professor at UC Berkeley
  - Author of Artificial Intelligence: A Modern Approach (most used text book in AI)
- General audience book (no technical background required)
- Published: October 8th, 2019

Gaining access to considerably greater intelligence would be the biggest event in human history

Everything civilization has to offer is the product of our intelligence

Might be the last in human history

How to make sure that it is not

The New York Times

Opinion

How to Stop Superhuman A.I. Before It Stops Us

The answer is to design artificial intelligence that's beneficial, not just smart.

By Stuart Russell

Dr. Russell is a professor of computer science at the University of California, Berkeley.

Oct. 8, 2019

TC

Loading...

Search Q

Disrupt Berlin 2019

Startups

Apps

Gadgets

Videos

Audio

Extra Crunch

Newsletters

Events

Advertise

Crunchbase

More

Extra Crunch

An interview with Dr. Stuart Russell, author of 'Human Compatible, Artificial Intelligence and the Problem of Control'

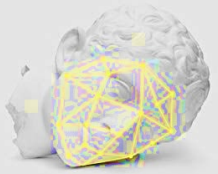
Ned Desmond @neddesmond

Apple

Image Credits: vasabii / Getty Images

Stuart Russell

HUMAN COMPATIBLE



AI and the Problem of Control

allen lane

BBC Sign in Home News Sport Reel Worklife Travel Future Culture More

SOUNDS

HARDtalk

Professor of Computer Science at University of California, Berkeley - Stuart Russell

The Telegraph News Politics Sport Business Money Opinion Tech Life & Style Travel Culture

Gadgets Innovation Big tech Start-ups Politics of tech Gaming

Premium

Technology Intelligence

Meet the man with the 'off-switch' for when the robots come for us

share Save 3

Professor Stuart Russell thinks he has worked out how to stop robots getting out of control

Follow By Harry de Quetteville

1 OCTOBER 2019 • 6:00AM

Nearly 10 years ago Stuart Russell, one of the world's leading experts on artificial intelligence, was on the Paris Metro in

FOLLOW THE T

Follow on Facebo

Follow on Instag

Support The Guardian

Available for everyone, funded by readers

Contribute Subscribe

Search jobs Sign in Search The Guardian International edition

News Opinion Sport Culture Lifestyle More

Books Music TV & radio Art & design Film Games Classical Stage

Book of the day

Science and nature books

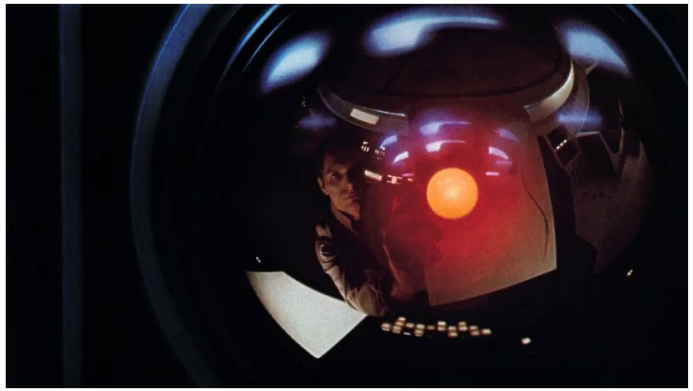
Human Compatible by Stuart Russell review – AI and our future

Ian Sample

@iansample

Thu 24 Oct 2019 07.31 BST

56

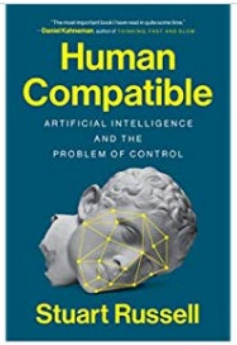


▲ Hal, the computer in *2001: A Space Odyssey* (1968), studies astronaut Dave Bowman, played by Keir Dullea. Photograph: Cinetext Bildarchiv/Allstar/MGM

Creating machines smarter than us could be the biggest event in human history – and the last

Elon Musk @elonmusk

Worth reading "Human Compatible" by Stuart Russell (he's great!) about future AI risks & solutions



Prime ★★★★★ (16 Reviews)

Human Compatible: Artificial Intelligence and the Problem of Control

Agenda

- Part I: Intelligence in humans and machines
- Part II: The problem of control
- Part III: A new approach to artificial intelligence

- Legend:
  - Text in black: Stuart Russel
  - Text in blue: My (≈AiTech, I believe) point of view + other authors

My point of view

1. Human-AI alignment is **not** a purely technical research problem

2. AI interacting with humans becomes a **complex socio-technical system**

3. I prefer to discuss **meaningful human control** over Narrow AI than Superintelligent AI

Meaningful Human Control

Humans not computers and their algorithms should ultimately remain in control of, and thus **morally responsible** for relevant decisions

Tracking condition

…respond to the relevant moral reasons of the relevant humans and the relevant factors in the environment in which the system operates…

Tracing condition

…possibility to always trace back the outcome of its operations to at least one human along the chain of design and operation…

Santoni de Sio, F., van den Hoven, J, 2018. Meaningful human control over autonomous systems: A philosophical account. **Frontiers in Robotics and AI** 5 (15).

Part I

Intelligence in humans and machines

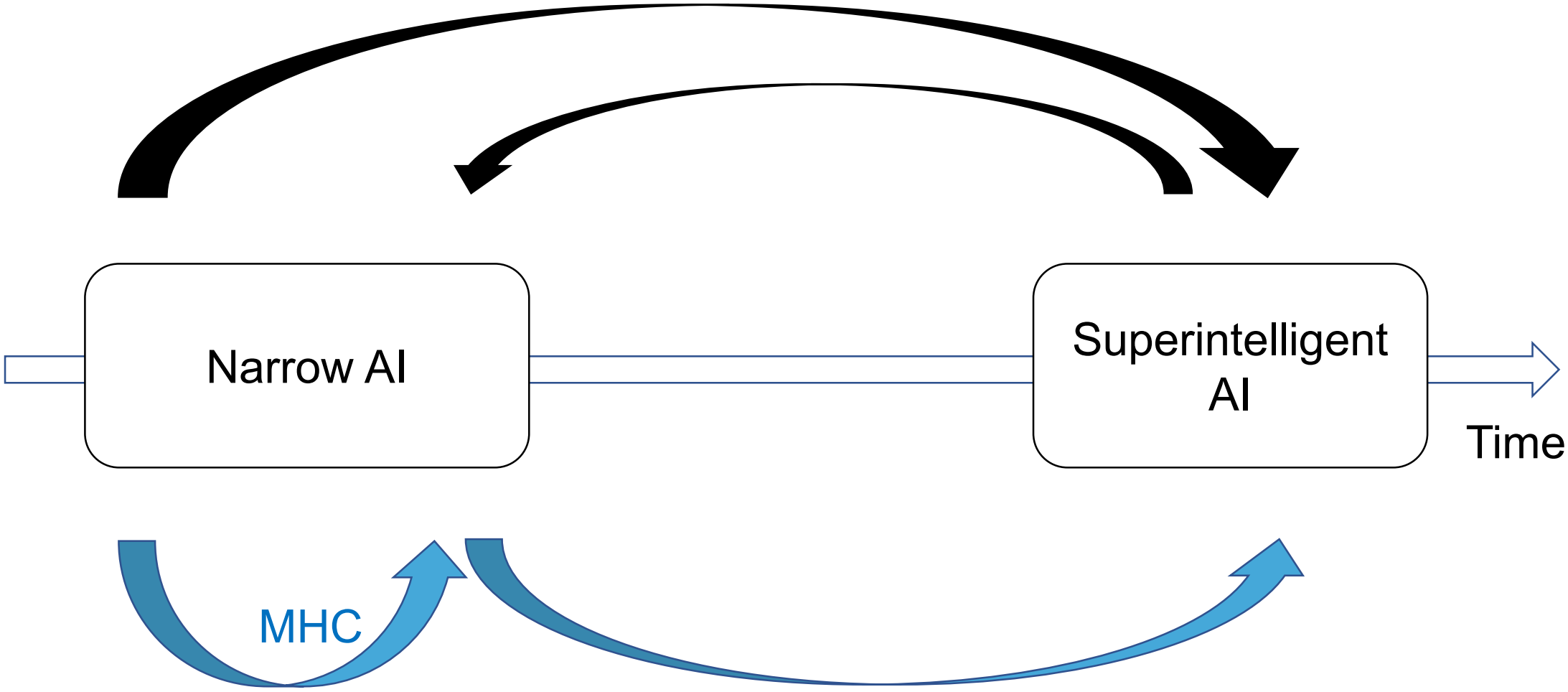
Introduction

- Are we about to be overtaken by superintelligent machines?
  - Probably not, but…
- *Even if "AI rise" is quite unlikely, we should prepare for it*

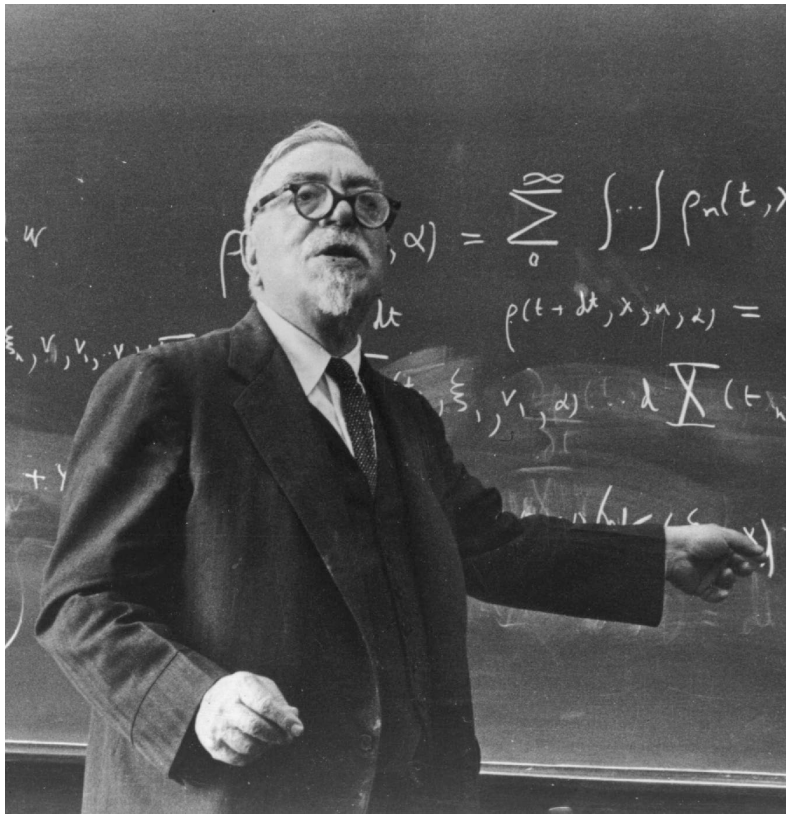
- Russel focus on superintelligent AI (≈AGI) (as Nick Bostrom, Max Tegmark, researchers at OpenAI, Deepmind, MRI, among many others)

- Problems of control, societal and ethical impacts, are and will increasingly affect us long before or even if we never reach AGI

Introduction



What went wrong?

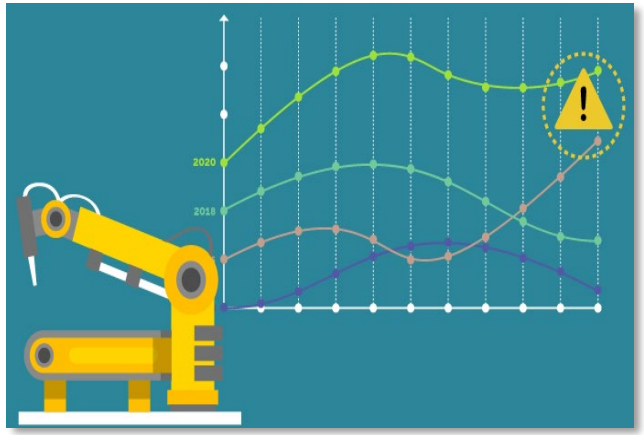
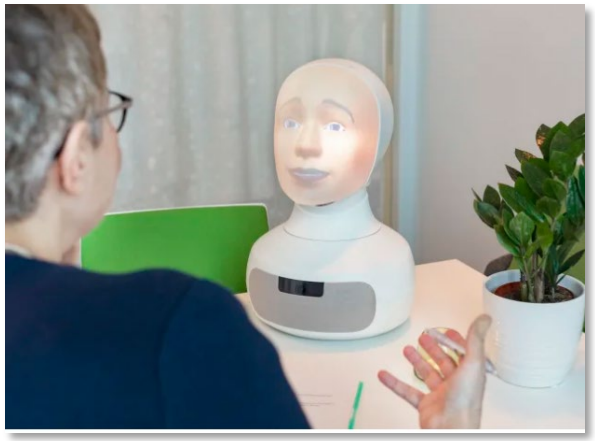


"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere (…), then we had better be quite sure that the ***purpose put into the machine*** *is the* ***purpose which we really desire****…"*

Wiener, Norbert. Some moral and technical consequences of automation. Science, 131, 1960.

Part II

The problem of control

	Predictive maintenance	Social media	Job suitability prediction
			
Criterion	Repair costs	Relevance of content	Avoid hiring unsuitable
Features	Vibration spectrum Resource usage Current signature	Vocabulary Social interactions Web presence	Facial emotion Voice timbre Vocabulary
Optimiser	Maintenance planning	Engagement	High volume selection
Optimised	Reduced value of expertise	Political polarization	Reduced self presentation

*This slide was based on ideas and previous presentations from Catholijn Jonker and Inald Lagendijk

The problem of control

- The Gorilla Problem
  - Can humans maintain their **supremacy** and **autonomy** in a world that includes machines with **greater** intelligence?

- The Humans Right Problem
  - Can humans maintain their **fundamental rights** in a world that includes machines with **great** intelligence/computational power/adaptability?

The problem of control

- The Kind Midas problem
  - Legendary king in ancient Greek mythology

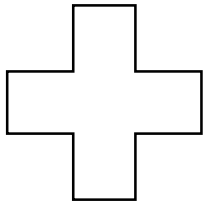
- *“We may suffer from a failure of value alignment”*
- Until recently:
  - Limited capabilities of AI → Limited ìmpact in the world
  - Now: Optimizing X Optimized

Part III

A new approach to artificial intelligence

A new approach

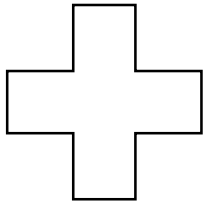
- Traditional approach to AI:
  - *"Machines are intelligent to the extent that their actions can be expected to achieve **their objectives**"*



Optimizing machines

Also on control systems, economics, operation research …

- Revised approach to AI:
  - *"Machines are **beneficial** (and intelligent) to the extent that **their** actions can be expected to achieve **our** objectives"*



Uncertainty about what our objectives are (as a feature, not a bug)

Social sciences

Principles for Beneficial Machines

1. ***The machine's only objective is to maximize the realization of human preferences (altruistic machines)***
2. ***The machine is initially uncertain about what those preferences are (humble machines)***
3. ***The ultimate source of information about human preferences is human behavior (learning machines)***

- Principles as a guideline for AI researchers, not explicit laws for AI behavior
- Focus of the book: One human interacting with one machine

Principle 1 → Altruistic machines

1. ***The machine's only objective is to maximize the realization of human preferences***

- Preferences cover everything one might care about in the future
- **Assumption: *An adult human has roughly consistent preferences over future lives***
- Rationality
  - *"Maximizing expected utility may not require calculation… purely external description"*
  - *"We are much further from being rational than a slug is from overtaking the starship enterprise traveling at warp nine"*

Principle 1 → Altruistic machines

David Leslie
Alan Turing Institute

"… he ignores the strain of twentieth-century thinking whose **holistic**, **contextual understanding of reasoning** has led to a humble acknowledgement of the existential limitations of intelligence itself. As a consequence, Russell ultimately falls prey to the **techno-solutionist idea that intelligence can be treated as an 'engineering problem'**, rather than a constraining dimension of the **human condition that demands continuous, critical self-reflection**"

Leslie, D. (2019). Raging robots, hapless humans: the AI dystopia. *Nature*, *574*(7776), 32.

Principle 1 → Altruistic machines

- **Assumption: A*n adult human has roughly consistent preferences over future lives*,** but…
  - Machines modify human preferences (by modifying human experiences) because this makes easier to satisfy one's preferences (or a given utility function)

- "*First, we shape our buildings and then our buildings shape us*"
  Winston Churchill

Principle 2 → Humble machines

2. The machine is initially uncertain about what those preferences are

- Creates a positive incentive for a machine to allow itself to be switched off (or, to ask for help / guidance / support)
- Uncertainty is a key concept on modern AI (> 1980s), but it was mostly ignored in the objective functions
- Moral uncertainty

Principle 3 → Learning machines

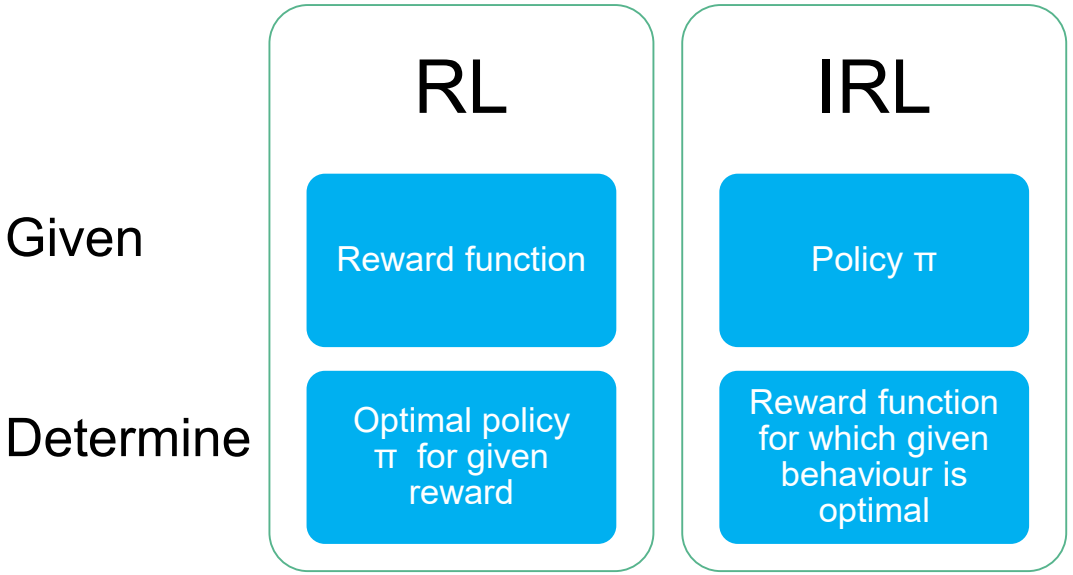
3. The ultimate source of information about human preferences is human behavior

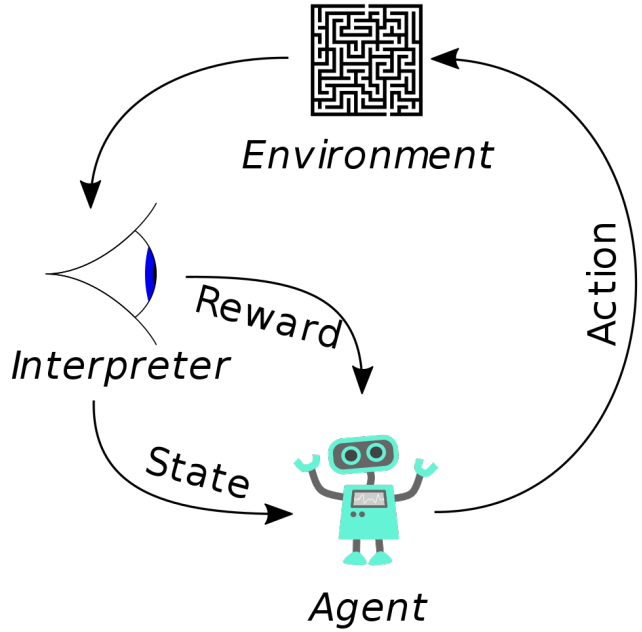
- Provides a grounding for what is meant by human preferences
- Preferences rather than human values: *"Values might lead to a confusion that we want to put our own values to a system, in other words, preconceptions about morality"*
- Preconceptions on morally acceptable behavior are relevant for MHC
- For Meaningful Human Control over (Narrow) AI, **values** better represent human's *relevant moral reasons*[1]
    - Abstract
    - Context independent
  - Humans use values and norms in folk explanations of their behavior[2]

[1]Dignum, V. (2019). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer International Publishing.
[2]Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Principle 3 → Learning machines

- **Inverse reinforcement learning (IRL)**: Given measurement of an agent's behaviour over time, in a variety of circumstances, determine the reward function being optimized*





*Ng, A., Russel, S, 2000. Algorithms for inverse reinforcement learning. **In: International Conference on Machine Learning (ICML)**.

Principle 3 → Learning machines

- *“A robot has to understand something about the cognitive processes that generate its behavior”*
  - Humans have an advantage: we use our own mind as “simulator”
- We need to work together with social (and natural) sciences
- Even with a simplicity prior, it is not possible to, simultaneously, estimate one’s preference and their rationality
  - We need simple “normative” assumptions, which cannot be deduce exclusively from observation
- Combine cognitive models with machine learning[1]
  - Train neural networks with synthetic data generated by cognitive models

[1]Peterson, J., Bourgin, D., Reichman, D., Griffiths, T., Russel, S. "Cognitive model priors for predicting human decisions." International Conference on Machine Learning (ICML), 2019.
[2]ARMSTRONG, Stuart; MINDERMANN, Sören. Occam's razor is insufficient to infer the preferences of irrational agents. In: Advances in Neural Information Processing Systems. 2018.

Principle 3 → Learning machines

- Should a machine act to support one's preferences in all situations?
- *"Machines may need to treat differently those who actively prefer the suffering of others"*

- Necessity to root on societal (agreed) ethical principles

- Is-ought problem / Naturalistic fallacy:
  - No ought-judgment may be correctly inferred from a set of premises expressed only in terms of 'is


PRINCIPIA ETHICA

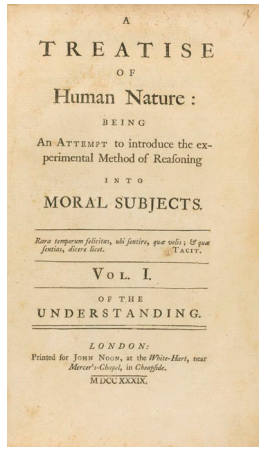
BY

GEORGE EDWARD MOORE

"Everything is what it is, and not another thing"
BISHOP BUTLER

CAMBRIDGE
AT THE UNIVERSITY PRESS
1903

Georgoe Moore
Principia Ethica (1903)


A
TREATISE
OF
Human Nature:
BEING
An ATTEMPT to introduce the experimental Method of Reasoning
INTO
MORAL SUBJECTS.

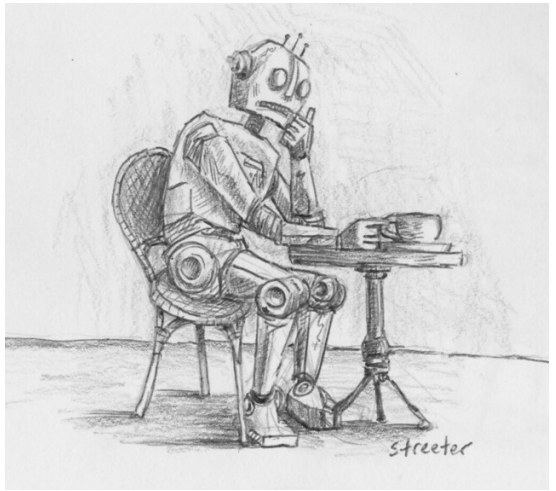
Rara temporum felicitate, ubi sentire, quæ velis; & quæ sentias, dicere licet.
TACIT.

VOL. I.
OF THE
UNDERSTANDING.

LONDON:
Printed for JOHN NOON, at the White-Hart, near Mercer's-Chapel, in Cheapside.
MDCCXXXIX.

David Humme
A treatise of Human Nature (1739)

Principle 3 → Learning machines



- Assistance games:
  - *“We don’t want the robot to want coffee!”*
  - Cooperative Inverse Reinforcement Learning:
    - “Robot” + Human
    - Partial information
      - Human knows the reward function
      - Robot’s payoff is exactly the human’s actual reward
    - Solutions may involve active instruction by the human and active learning by the robot

*Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In Advances in neural information processing systems (pp. 3909-3917).

Final remarks

Final remarks

- Altruistic and learning machines
  - Technical methods are useful to cope with system speed and information processing
  - Social (and natural) sciences are important to understand the complexity and mechanisms of human decision-making and morality
- Humble machines
  - Adaptation/tracking
  - Should be rooted on ethical principles and social norms
- Beyond one human and one robot
  - Assistance games
  - How to define trade-offs? Ethics, social choice, psychology

Final remarks

- *"The real control problem isn't managing the coming of transcendent superintelligent creatures. More critically,* ***it has to do with reining in the triumphalist creators who may be developing increasingly "autonomous" AI technologies under the auspices of the misguided definition of intelligence"****

*Leslie, D. (2019). Raging robots, hapless humans: the AI dystopia. *Nature*, *574*(7776), 32.

HUMAN COMPATIBLE

AI and the Problem of Control

A critical discussion by
Luciano Cavalcante Siebert