On the role of ethics in (meta)cognition

Vincent C. Müller

TU Eindhoven (U Leeds / Alan Turing Institute)
http://www.sophia.de

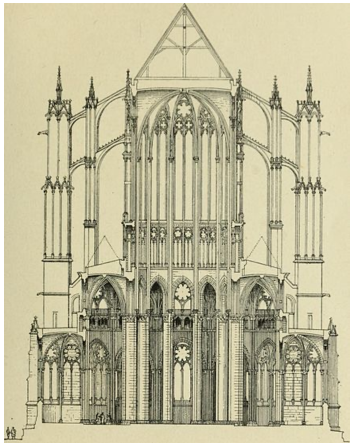
14.01.2022

Outline

1. Ethics: Normative vs. descriptive
2. Singularity & Orthogonality → Existential Risk?
3. The trick with frames
4. As-if goals and goals
5. "What makes us so successful?" … normative metacognition

1) Biographical Background: "Ethics of AI & Robotics"

(Stanford Encyclopedia of Philosophy - https://plato.stanford.edu/entries/ethics-ai/)



1. Introduction

2. Main Debates

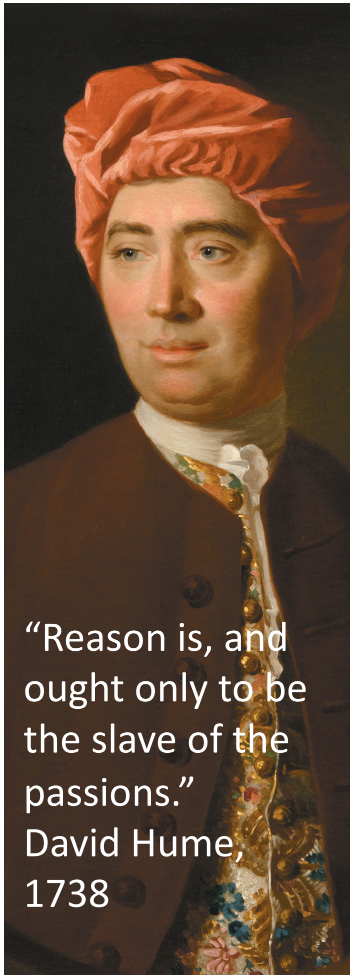
2.1-2 Data: Privacy & Manipulation

2.3-4 Epistemology: Opacity & Bias

2.5-7 Robot Ethics: Automation, Interaction, Autonomy

2.8-9 Concepts (Agency, Responsibility, Autonomy …)

2.10 Singularity (Superintelligence)



"Reason is, and ought only to be the slave of the passions."
David Hume, 1738

Descriptive vs. Normative

- We now have an unprecedented amount of data.
- Data is likely to change the world.
- People have values and follow norms.

- An engineer should act in the public interest.
- Survival is more important than privacy.
- You should have told me that my girlfriend is cheating on me.

“Was kann ich wissen?
Was soll ich tun?
Was darf ich hoffen?”
I. Kant, 1781

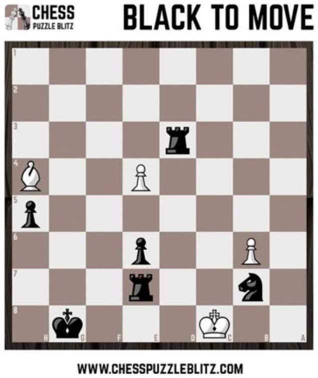
Ethics

- Etymology:
  “mores”, “ethos” (ἤθος”) = customs, character, ‘what one does’

- Ethics:
  systematic reflection on the normative

- Basic question of ethics:
  “What should I do?” (I. Kant)
  - What should be my next action?
    (Ethics is part of an advanced theory of rational choice.)
  - Traditionally: ethics ⊂ normative

- → Ethics is already part of what we are doing



2.
Singularity & Orthogonality → Existential Risk?

a) Superintelligent AI is a realistic prospect and it would be out of human control (Singularity claim)

b) Any level of intelligence can go with any goals (Orthogonality thesis)

XRISK

→ Superintelligent AI poses an existential risk for humanity

Müller, Vincent C. and Cannon, Michael (2021), 'Existential risk from AI and orthogonality: Can we have it both ways?', *Ratio*, 00, 1-12. https://onlinelibrary.wiley.com/doi/10.1111/rati.12320

3. The trick with frames

- The Argument from Superintelligence to Existential Risk assumes general intelligence
  - G-Intelligence = "like us" = "general cognitive ability" with intentions to achieve goals

- Orthogonality thesis assumes instrumental intelligence
  - I-Intelligence = instrumental intelligence to reach given goals ('utility functions')

	Orthogonality	Existential Risk
Instrumental Intelligence	Consistent	Inconsistent?
General Intelligence	Inconsistent?	Consistent

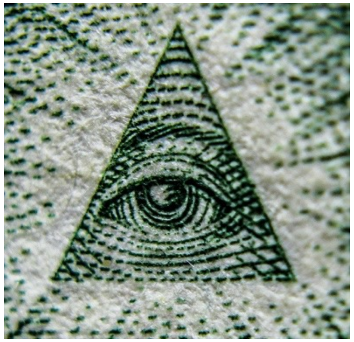
Intelligence[i]

τέλος

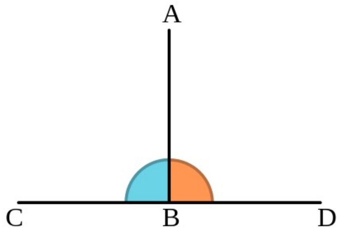
- Instrumental: "For our purposes, 'intelligence' will be roughly taken to correspond to the capacity for instrumental reasoning […]. Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal." (Bostrom 2012, 73)
- "By "intelligence", we here mean something like skill at prediction, planning, and *means-end reasoning* in general. This sense of *instrumental cognitive efficaciousness* is most relevant when we are seeking to understand what the causal impact of a machine superintelligence might be." (Bostrom 2014: 107)
- "Intelligence measures an agent's ability to achieve goals in a wide range of environments." (Legg and Hutter 2007: 402).
- Problem-solving
- Decision theory: rational = maximises expected utility

Intelligence^g - “like us”

- General intelligence involves metacognition
  - We can widen the frame of reflection (Should I try to win this game of chess?)
  - We can reflect on goals (Is winning important?)
- Problem-solving AND Problem-defining



Summary: We can’t have it both ways



- The singularity claim[g] and the orthogonality thesis[i] might both be true … but only with those indices
- There are forms of intelligence that make Xrisk plausible, and others that make orthogonality plausible
- We see no interpretation of “superintelligence” that allows for both Xrisk & orthogonality

XRISK

	Orthogonality	Existential Risk
Instrumental Intelligence	Consistent	Inconsistent?
General Intelligence	Inconsistent?	Consistent

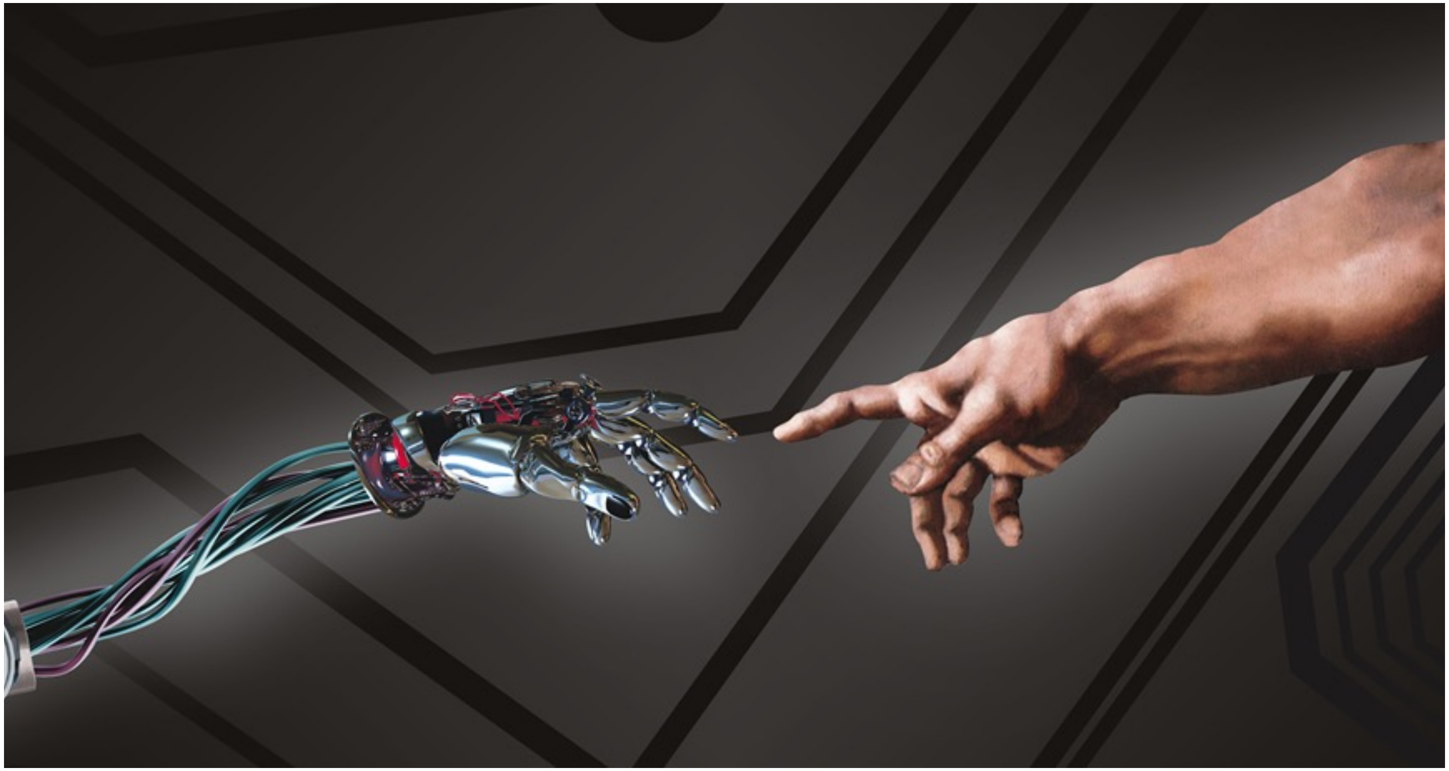
4. As-If Goals & Goals

- AI “goals”
  - Autonomous car turns right at traffic light
  - Chess computer is selecting a move
  - Robot returns to charging station
- “Goal" as externally defined setpoint
  - 2nd order goals (derived from human goals) are useful for “adaptive control”
- AI: reflex agents, goal-based agents, utility-based agents
  - „In short, *a rational agent acts so as to maximise expected utility*. It’s hard to overstate the importance of this conclusion. In many ways, artificial intelligence has been mainly about working out the details of how to build rational machines.“
  - “AI has adopted the standard model: we build optimising machines, we feed objectives into them, and off they go.” (Russell 2019, 172)

5. "What makes us so successful?" – Normative metacognition!

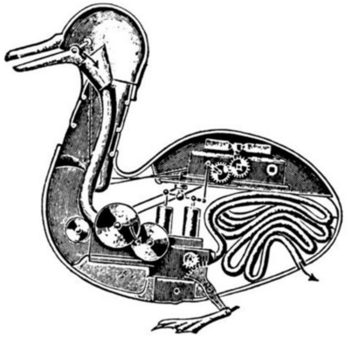
- **Instrumental intelligence**
  - Rational choice of action that has the highest subjective expected utility
- **Embodied Intelligence**
  - Human bodies are part of the cognitive system & the explanation …
- **Extended & Embedded Intelligence**
  - Human culture & artefacts are part of the cognitive system & the explanation

- **General intelligence with metacognition**
  - "It is an inherent property of intelligence that it can jump out of the task which it is performing, and survey what it has done; it is always looking for, and often finding, patterns." (Hofstadter 1979: 37).
  - Widening the frame, defining problems
  - Having real goals (with subjective value)
  - Ability to reflect on goals, i.e. normative metacognition (also called "ethics")



Thank You!

1. AI vs. As-If AI (AI-AI)



- "If it looks like a duck, walks like a duck and quacks like a duck, then it's a duck" (vs. de Vaucanson?)

vs.

- We always have to distinguish between really being a duck and just behaving as-if (a correct input-output function)

- Current AI does not "really" have moral agency, intelligence, goals, etc. etc. … the old muppets say

Real Goals

τέλος

- Has subjective value for the agent (≠ “utility function”, ≠ intrinsic value) &
- Agent aims for it &
- Agent is aware of goal (≠ end)

- Goals can be the result of choice or of natural (teleological) processes
- In folk psychology, goals (desires) + beliefs explain action