Artificial Intelligence, Understanding, and Moral Characters

Mahdi Khalili

|PhD in Philosophy, Vrije Universiteit Amsterdam

|Postdoc, Institute for Research in Fundamental Sciences, Tehran

|Researcher, Ethics Department of Cyberspace Research Center, Tehran

Introduction

- Unintelligibility of AI systems that employ ML techniques in DNN
- My approach: integrated philosophy and ethics of science and technology
  - Claus Beisbart and Tim Räz (2022): Four tasks for philosophers of science
- Which concept?
  - Transparency (≠ Opacity)
  - Explainability
  - Interpretability
  - **Understandability (≠ Unintelligibility)**
- Structure of my presentation:
  - An Argument for Understanding
  - Explanation may be valuable in use contexts
  - Qualitative understanding with a human face

An Argument for Understanding

1. Unintelligible artificial systems foster characters that are indifferent to the understanding of reasons (for actions and decisions).
2. Characters indifferent to understanding do not realize their (moral) capacity.

Thus, in the techno-social context of unintelligible systems human (moral) capacity can hardly be realized.

An Argument for Understanding

1. **Unintelligible artificial systems foster characters that are indifferent to the understanding of reasons (for actions and decisions).**
2. Characters indifferent to understanding do not realize their (moral) capacity.

Thus, in the techno-social context of unintelligible systems human capacity can hardly be realized.

- **Normativity of technologies**
  - **Hans Radder (2019, chapter 2):** For a type of technology to function stably and reproducibly, the relevant techno-social context *should* be appropriate
  - I: Human characters indifferent to reasons are among the conditions of the realization of unintelligible (AI) systems
  - **Technologies shape characters**

An Argument for Understanding

1. Unintelligible artificial systems foster characters that are indifferent to the understanding of reasons (for actions and decisions).
2. **Characters indifferent to understanding do not realize their (moral) capacity.**

Thus, in the techno-social context of unintelligible systems human capacity can hardly be realized.

- Moral perspective: "a reliable disposition to *attend to, discern, and understand moral phenomena as meaningful parts of a moral whole*" (Shannon Vallor 2016, p. 149)
- **Understanding → Moral perspective → Practical wisdom**

An Argument for Understanding

1. Unintelligible artificial systems foster characters that are indifferent to the understanding of reasons (for actions and decisions).
2. Characters indifferent to understanding do not realize their (moral) capacity.

Thus, in the techno-social context of unintelligible systems human capacity can hardly be realized.

- **The contraction of the "space of reason"**
- **The contraction of the moral extent of characters deploying unintelligible systems**

Explanation May Be Valuable in Use Contexts

- Nathan Colaner (2021): Unexplainable systems are dehumanizing. People are not able to fully actualize themselves unless they are able to meaningfully participate in the decision-making procedure.
- But
  - Explanations are not intrinsically valuable
  - XAI systems *themselves* are neither sufficient nor necessary for good actions or decisions.
- My argument is not against unintelligible AI systems themselves, but rather against employing unintelligible AI systems in contexts in which morally relevant factors rely on our understanding of these systems.

Against the intentional stance

- John Zerilli et al. (2019): a justification of DNN should be provided at the level that Daniel Dennett (1987) calls the “intentional stance”:
- But
  - The attribution of this stance to machines is either metaphorical or metaphysically problematic.
  - And it is confusing to consider the different explanation styles employ beliefs etc.

      - > If 10% or less of your driving took place at night, you would have qualified for the cheapest tier.
      - > If your average miles per month were 700 or less, you would have qualified for the cheapest tier.
      - (Binns et al. 2018, p. 6)
  - The attribution of the intentional stance to machines is acceptable only metaphorically
  - An explanation can be stated by technical, scientific terms.

Qualitative Understanding with a Human Face

My alternative: Henk de Regt's (2017) theory of scientific understanding

- XAI could draw inspiration from how scientists make unintelligible phenomena/models understandable
- Three relevant elements in his view:
  - Understanding is a **human** capacity
  - Understanding is **context-dependent**
  - Understanding is **qualitative**

"A scientific theory T is intelligible for scientists (in context C) if they can recognize **qualitatively** characteristic **consequences** of T **without performing exact calculations**." (De Regt, 2017, p. 102)

- Qualitative: *Computational* processes that take place at the level of the architectural innards of the system
- Consequences include
  - Qualitative sense of how the system produces its outputs
  - Recognition of moral consequences → **prudential judgment**

Conclusion

I have employed an **integrated philosophy and ethics of science and technology** approach to argue against unintelligible artificial systems

Unintelligible artificial systems foster characters that are indifferent about the reasons of actions and decisions; these characters cannot realize their moral capacity; therefore**, in the context of unintelligible systems human capacity can hardly be realized**.

XAI techniques make AI systems intelligible if they can provide **humans** with **qualitative consequences** of the AI system that are relevant to its **context** of use.