Unintelligible Artificial Intelligence and Virtue Ethics

Mahdi Khalili, Institute for Research in Fundamental Sciences (Iran)

International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: understanding, artificial Intelligence, normativity of technology, virtue Ethics

Extended Abstract

Unintelligible artificially intelligent systems produce outputs that even their makers do not understand why specific patterns have been extracted from a given dataset. In this paper, I explain my concern about unintelligible technologies (including unintelligible AI systems), which can be set out in the form of an argument as follows.

> (1) Unintelligible technologies have the potential to require that agents be indifferent to the understanding of reasons (for actions and decisions).
>
> (2) Agents who are indifferent to the understanding of reasons (for actions and decisions) do not realize their moral capacity.

Therefore, unintelligible technologies have the potential to require that agents do not realize their moral capacity.

The first premise accepts the normativity of technology. That is, for a functional system to work in a stable and reproducible manner, certain social and technical contexts should be established. I follow Hans Radder (2019, chapter 2) on the definition of technology and its inherent normativity. AI systems are kinds of technology, so they are normative as well, and our socio-technical conditions must be appropriately changed to fit them if those systems are to realize their full potential. My claim is that human agents who are indifferent to the reasons for decisions and actions are among the conditions for the full realization of unintelligible AI systems. Humans typically ask for the reasons for decisions and actions. In particular, they do so regarding technologies that impact their lives, including AI systems with their influential applications in a variety of domains. The typical questions raised by humans interested in understanding unintelligible AI systems disturb the proper functioning of these systems. On the other hand, agents with "indifferent characters" are among the necessary conditions for these systems to function effectively, and in this sense unintelligible AI systems have the potential to require that agents be indifferent to the realm of reason; such agents can enjoy the efficiency of unintelligible AI systems without going to the trouble of asking unanswerable questions about them. As a result, even if AI systems do not result in explicitly immoral consequences, they can bring about an "indifferent" society, whose members do not usually, or are usually unwilling to, ask (moral and political) questions.

The second premise explains why the constitution of agents with characters indifferent to understanding is undesirable. Such agents do not realize their capacity of practical wisdom, or what Shannon Vallor (2016, chapter 6) calls "technomoral wisdom", in which "technomoral virtues" – i.e., virtues that are necessary to live a good life in the age of emerging technologies – are integrated. One of these technomoral virtues is "moral perspective", which Vallor defines "as a reliable disposition to attend to, discern, and understand moral phenomena as meaningful parts of a moral whole" (2016, p. 149). A person indifferent to

understanding reasons for decisions and actions does not discern or understand moral phenomena appropriately. This person neither pays serious attention to morally relevant factors, nor grasps the importance of these factors in the broader context of a decision or an action. Indeed, the moral perspective explains the connection of these two human capacities: understanding and practical wisdom. It also clarifies the key role of the former in the latter. Moreover, because the moral perspective is "an essential disposition of a virtuous person" (2016, pp. 149–150), those with an insufficient moral perspective cannot practice other virtues such as justice, honesty, care, and civility. As a result, those who lack adequate understanding cannot cultivate (technomoral) virtues.

The logical result of both premises is that unintelligible technologies (including unintelligible AI systems) have the potential to require that agents not realize their moral capacity. An implication of this conclusion is that the demand for understandability of artificially intelligent systems signifies a struggle for virtuous characters and communities. Therefore, in order to have a society with good personalities, we should avoid unintelligible AI systems and take steps to design understandable ones.

In the remainder of the paper, I draw on the literature on scientific understanding to suggest that an artificially intelligent system can be rendered understandable if a _qualitative_ account of the consequences of its use in context is provided. I suggest that the project of developing XAI methods could draw inspiration from how scientists make unintelligible phenomena or models understandable. There are several theories/models in natural science that are predictively successful, although they are faced with the "black box" problem. Scientists usually formulate models to make these black boxes understandable. I see the project of developing and using XAI methods as being in a similar vein. XAI developers desire to make unintelligible AI systems understandable by constructing simple intelligible models.

There are several theories of scientific understanding. Among them, I refer to Henk De Regt (2017), whose Criterion for the Intelligibility of Theories follows:

> CIT: A scientific theory T (in one or more of its representations) 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)

How can CIT help us to establish a criterion for the understandability of artificial intelligence? To answer this question, we should explore what it means to possess a qualitative recognition of an AI system without having exact calculations. I suggest that an AI system, and its predictions and consequences, can be recognized qualitatively when our recognition does not depend on the computational processes that take place at the level of the architectural innards of the system. Causal reasoning, visual representations of significant mechanisms, and discovering continuity/resemblance between the AI system and other understandable systems provide kinds of qualitative understanding, but there may be several other conceptual tools. Intelligibility is a pragmatic and context-dependent property, so the achievement of the intelligibility of an AI system is related to the characteristics of the system, its contexts of use, the skills of XAI developers, and the stakeholder(s) to whom the system should be intelligible (see also Zednik 2021).

According to CIT, agents "can recognize qualitatively characteristic consequences of T without performing exact calculations". What might be the nature of the qualitative consequences of an AI system? The answer depends on the agent to whom the system should be understandable. For instance, AI scientists and developers should possess some qualitative sense of how the system produces its outputs, or the users should understand how the system affects their futures. In this regard, I would like to highlight the point that the consequences are not merely epistemological, but are moral as well. Virtuous characters possess prudential judgment, "the cultivated ability to deliberate and choose well, in particular situations, among the most appropriate and effective means available for achieving a noble or good end" (Vallor 2016, p. 105). Although virtue ethics is in tension with merely consequentialist normative ethics, having the ability in prudential judgment requires being able to consider some foreseeable consequences of a decision or an action. Prudent characters examine the moral consequences of the AI systems they use to see if these systems are appropriate tools to achieve good purposes. As a result, prudent characters need to understand AI systems in the sense that they should be able to recognize the morally relevant, qualitatively characteristic consequences of the systems without having knowledge of complex computations.

The novelty of this paper consists, first, in its approach. The understandability of artificial intelligence is usually studied either from the perspective of philosophy of science (and epistemology, more broadly) or from an ethical perspective. This paper attempts to maintain both these perspectives at the same time. Second, the argument of this paper against unintelligible technologies is novel. It focuses on the impact of using unintelligible technologies on their users' moral characters. The third novelty of the paper concerns its suggestion about the kind of understanding that should be provided by the methods that render AI understandable. In this regard, I draw on discussions of scientific understanding.

References

De Regt, Henk. (2017). Understanding scientific understanding. Oxford: Oxford University Press.

Radder, Hans. (2019). From commodification to the common good: Reconstructing science, technology, and society. Pittsburgh: University of Pittsburgh Press.

Vallor, Shannon. (2016). Technology and the virtues: A philosophical guide to a future worth wanting. Oxford: Oxford University Press.

Zednik, Carlos. (2021) Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. Philosophy & Technology, 34, 265–288.