Coeckelbergh, M. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. Sci Eng Ethics 26, 2051–2068 (2020). https://doi.org/10.1007/s11948-019-00146-8

Overview

As our lives are becoming more dependent on the presence of automated machines powered by artificial intelligence (AI), the problem of AI ethics has come to the surface.

This review analyses how a professor, Coekelbergh, M, at the Department of Philosophy from the University of Vienna, introduces a conceptual framework for responsible AI from the traditional Aristotelian criteria for responsibility; "The problem of responsibility attribution is initially approached by distinguishing between two Aristotelian conditions of responsibility, one is about control and a need to identify the responsible agent, and another which examines the agent's knowledge" (p. 2052).

He begins with the 'control' condition of the traditional criteria to define responsible agents and moves to the second condition, 'epistemic', to define responsible users(patients). The author then points out the limitations that come from the socially and structurally intertwined nature of AI systems and introduces the concept of responsible relation between the responsible agents and responsible patients. The author claims that this relation is the key to build a responsible AI and to enhance answerability as explainability.

1. Who or what can be responsible agents - 'control' condition

The chapter (The Problem of Responsibility Attribution: Who or What is the Agent of Responsibility) focuses on the 'control' condition to attribute responsibility. This condition defines responsible agents as having sufficient control over their actions and decisions. The author claims these decisions and actions should be made voluntarily and consciously. Thus he restricts the agents of responsibility to humans because AI does not have free will nor know what it does; "the remainder of this paper will assume that AI technologies do not meet traditional criteria for a full moral agency such as freedom and consciousness and that therefore they also cannot be (held) responsible" (p. 2054). However, delegating responsibility to humans brings about other issues emanating from the discrepancy between humans and machines.

First, the author says there can be a 'responsibility gap' or 'problem of time', which implies that humans cannot always make an end decision because humans may not respond in time against automated decisions or human decision-making may be too slow. For example, many organizations or governmental entities have enforced AI ethics with the help of the concept of human-in-the-loop or -on-the-loop (for example, European Commission, Ethics Guidelines for Trustworthy AI, 2018), but when humans cannot intervene with the end decision, how can humans be responsible agents?; "What if there is no human “in the loop”, and what if it doesn’t even make sense anymore to say that there should be humans “*on* the loop” (supervising) since they cannot intervene anyway?" (p. 2055).

Second, the actions AI makes are not caused by one single individual but rather by multiple stakeholders ('many hands') at different time dimensions ('temporal dimension'), which makes it difficult not only to identify individuals who are responsible for a particular action at a specific time but also to distribute responsibilities among those individuals. The author claims that 'traceability' is an important way to identify responsible agents in the series of causal chains of human agencies by allowing us to follow back the causal history; "If there is no record and if this cannot be traced, there is a responsibility attribution problem. This is why traceability is an important way to operationalize responsibility and explainability" (p. 2057).

Third, AI systems consist of various technological components ('many things') that causally contribute to the technological interaction. Each component is intertwined with one another and associated with humans involved in each component ('many hands'). Thus, this complicated link of components causes the same responsibility attribution problem from 'many hands'. The author then defines the responsibility for technology as a matter of interfaces; "In this sense, responsibility for technology is not only a matter of faces but also of interfaces" (p. 2058).

2. Responsible use of AI - 'epistemic' condition

While the 'control' condition was to define responsible agents, the 'epistemic' condition is to define a responsible use of AI more from the end-user side; "A man may be ignorant, then, of who he is, what he is doing, what or whom he is acting on, and sometimes also what (e.g., the instrument) he is doing it with, and to what end (e.g. for safety), and how he is doing it" (p. 2058). In the chapter (Knowledge Problems: Transparency and Examinability), the author claims that the ignorance of humans can emerge in two ways in the context of AI.

First, AI applications are usually fueled by data, which is a collection of pervasive human bias and disparate factors in society over history, thus leading humans to ignorance; "technology is always more-than-instrumental and is used in a social and relational context. But usually neither of developers nor users of technology are aware of this" (p. 2059).

Second, the black-box nature of AI systems does not clearly explain how they arrived at their decision and recommendation. The author claims that this ignorance raises ethical questions. Thus he puts an extra emphasis on the importance of including a 'to whom' condition to stipulate those who may be affected by the unintended consequences and on the importance of 'transparency'; The author says "Users of AI and indeed managers or regulators may not always know who else is or has been involved in the development and use of the technology, and what they have been doing and intending. This renders responsible use and responsibility attribution and distribution difficult" (p. 2060), and "ignorance also hits those who create and use AI technologies, for instance when these technologies arrive at their recommendations or decisions in an insufficiently transparent way" (p.2061).

3. Responsible relation and answerability

The chapter (Include Responsibility Patients, or Responsibility as Answerability: Towards a More Relational View) combines the concepts of 'control' and 'epistemic' in a way that one can separate the responsible agents and subjects, whom the responsible agents should be responsible for, while providing transparency in the context of the use of AI. To this end, the author introduces the conceptual framework, 'responsibility relation', on which 'responsible AI' can be built by defining the 'patients'; "there is not only an agent of responsibility but also a patient who is affected by the action of the agent and who demands the agent acts responsibly in the sense that she is expected and asked to give reasons for her action" (p. 2061). The author uses 'responsibility relation' and 'answerability' interchangeably, and claims this to be the purpose of the development of a responsible AI, on which 'responsible innovation' can be built in a way that enhances the answerability by means of technologies, legal measures, and discussions (p. 2063). The motivation for this framework is, one can say, to minimize what cannot be controlled and expected, which the author refers to as a ‘tragic dimension’, to foster the responsibility of AI by introducing a relation among those involved in the development and use of AI because data on which AI is trained involve not only the social but also the historical aspects that may not be able to be eradicated but that may be able to be partially corrected by humans.

4. Conclusion

The author explained the importance of responsible AI not only from a temporal aspect of the data, but also from an individual and collective perspective. He then introduced a relational framework as a basis for responsible AI. Thus, it can be said that the article has achieved its primary goal of justifying the need for a conceptual framework for responsible AI.

Its strength, also its weakness at the same time, is the fact that the author restricted the agents that can be held responsible to humans throughout the article. This restriction, on the one hand, makes it intuitive for readers to understand how responsibility within automated systems, especially in the context of AI-based systems, can be attributed through the responsible relation and what can be considered to achieve the goal of responsible AI. However, on the other hand, it can implicatively lead readers to have the idea that AI alone cannot be held responsible. Although many researchers have agreed that it is inevitable to put humans in the life cycle of AI systems, one should not neglect the fact that AI is a system that operates on the basis of complicated mathematic formulas and interconnected components. Indeed, there have been plenty of research to tackle the problem of AI ethics not only from the humanities aspects, but also from the statistical and technical aspects; "In response to the fact that artificially intelligent technologies gain more agency, it might be tempting to explore whether the technologies themselves can be responsible agents. In ethics of computing and robot ethics there have been extensive discussions about this topic and the related issue of moral agency" (p. 2054). Therefore, although the author successfully suggested a moral framework for AI, how stand-alone responsible AI can be achieved remains to be discussed.

Jaeyeop Chung
Technische Universität München

I have composed the paper independently on my own. I have not used any sources and aids other than those indicated. The work does not contain any text passages generated with AI support.

Jaeyeop Chung is a student from at Technische Universität München, studying for the master program of Data Engineering and Data Analytics. He is from South Korea and speaks Korean, English, and a little bit of German. He is interested in the business aspect of data science and data governance.