My Beautiful Paper About Stuff

Davide Rigamonti

June 2023

Todo list

1 Introduction

The historical model for determining the responsibility of a company or an individual that develops an AI system isn't valid when looking at autonomous systems. A new perspective is needed to effectively discern the full extent of human responsibility when a system is said to be able to "act on its own".

The main goal of this article is to underline the presence of a responsibility gap (Matthias 2004) when analyzing all kinds of autonomous systems (in particular those capable of "learning"), and to provide possible approaches and solutions in order to tackle the problems that may arise from applying classical frameworks of responsibility in current and future contexts that employ AI technology. The issue at the core of this paper has already been studied by many (Matthias 2004; Sio and Mecacci 2021; Coeckelbergh 2020; Novelli, Taddeo, and Floridi forthcoming), offering different points of view and valuable insight, however, to this day there are still a lot of questions that remain unanswered. This paper is not aiming to give definitive answers to the issue at hand, instead it will just focus on the presented thesis by providing argumentative evidence and relevant case-studies with a particular focus on LAWS (Lethal Autonomous Weapon Systems); to do so, in section 2 it will start by giving semi-formal definitions of *responsibility* (including some peculiar variations) and its applicability on non-autonomous systems following the classical framework, while section 3 is devoted to presenting, assessing and analyzing the opposite situation, highlighting the flaws of the previous approach; section 4 will focus on the possible long and short term consequences of erroneous handling of the presented issues, while section 5 will offer some plausible solutions to mitigate and possibly prevent those consequences; to conclude, section 6 is dedicated to the acknowledgement of eventual points of objection and section 7 will wrap up the paper.

insert refs

2 Analyzing Responsibility

The term *responsibility* is often used as a broad expression to convey the notion of "someone" (which can be a person, an institution, a corporation or more generally, an *agent*) having the duty of upholding certain expectations defined by another (or the very same) agent towards a given goal; however, it's important to acknowledge the presence of derived terms that represent different "flavors" of *responsibility*, in this section we focus on the specific expressions that are most relevant to this paper, nevertheless, it's important to note that numerous taxonomies can be formulated.

In the Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, *accountability* is defined as a term which refers to the idea that one is responsible for their actions and consequences, therefore they must be able to explain their aims, motivations, and reasons (Commission, Directorate-General for Com-

munications Networks, and Technology 2020 as cited in Novelli, Taddeo, and Floridi forthcoming); it is essential to note the presence of an autorithy in charge of supervising the conduct of the agent held accountable and thus, sitting closely to the definition of *answerability* (Nissenbaum 1996) in contrast with the notion of *moral responsibility*, which assumes an internal analysis against the very own moral values of the agent of interest. An important distinction that needs to be made is the difference between *active responsibility* and *passive responsibility* (Poel, van de and Royakkers 2011): while the first term is appropriate for addressing the continuous and preemptive effort that one must make to care about a certain goal while it is being attained, the second is applicable in the event that something undesirable has already happened instead and can be subdivided in *accountability*, *blameworthiness* (indirect responsibility) and *liability* (economical/legal responsibility) according to (Poel, van de and Royakkers 2011).

Another important aspect of responsibility is its context, as the notion of *role responsibility* suggests, in our daily life we partake in social roles which come with their respective responsibilities that may be conflicting; in particular, *professional responsibility*, is a particular type of role responsibility which concerns the professional life of an agent (Poel, van de and Royakkers 2011).

Feinberg identifies two main causes for an agent to be held morally blameworthy for any given harm: *causality* (the agent's actions contributed in causing the harm) and *faultiness* (the actions were intentionally harmful or a result of negligent behavior) (Feinberg 1985). Another viable interpretation is the aristotelian one, which specifies the *control condition* (intent and freedom of action) and the *epistemic condition* (awareness or non-ignorance) (Fischer and Ravizza 1998). It's easy to see how these frameworks fit quite well with most engineering tasks due to the fundamental instrumentality of the produced artifacts, for example .

insert example

3 The Responsibility Gap

In his seminal work, Andreas Matthias mentions the presence of a *responsibility gap* (Matthias 2004) in the process of ascribing responsibility to agents involved in the activity of creating and interacting with learning automata. First of all, it's important to understand in which ways autonomous learning systems differ from traditional engineering artifacts; in most cases a learning automaton is able to return outputs following a certain logic that wasn't directly coded by a programmer, but it was "learned" by providing examples of expected behavior, this particular approach makes it possible to achieve outstanding feats but at the same time it may raise some concerns on the actual *control* that any agent may have on the system.

As we have discussed in the previous section, *knowledge and control* of the agent over the system are fundamental prerequisites for ascribing responsibility (and even more so accountability) in the traditional frameworks; however, when facing a system that has the properties of a *black box* we cannot assume that anybody would be able to know its inner workings, not the creator of the automaton, let alone its potential users. In addition, even the *causality* conditions may not be satisfied if we consider what is known in the literature as the *problem of many hands* (Poel, van de, Fahlquist, et al. 2012; Coeckelbergh 2020; Nissenbaum 1996) and consists in the impossibility of generalizing the single agent responsibility to a collective entity, as an unfeasible unfolding of long and convoluted chains of events is needed to fully understand the contribution of each agent. The problem of many hands is not restricted to the learning automata case, but it is certainly amplified by the fact that AI models are usually composed of many different modules (*problem of many things* Coeckelbergh 2020) which are the product of contributions given by many different people with different skills, ideals and goals.

4 c

5 d

6 e

7 f

References

Feinberg, Joel (1985). "Sua Culpa". In: *Ethical Issues in the Use of Computers*. USA: Wadsworth Publ. Co., pp. 102–120. ISBN: 0534042570.

Nissenbaum, Helen (1996). "Accountability in a Computerized Society". In: *Science and Engineering Ethics* 2.1, pp. 25–42. DOI: 10.1007/bf02639315.

Fischer, John Martin and Mark Ravizza (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Ed. by Mark Ravizza. New York: Cambridge University Press.

Matthias, Andreas (2004). "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata". In: *Ethics and Information Technology* 6.3, pp. 175–183. DOI: 10.1007/s10676-004-3422-1.

Poel, van de, I.R. and L.M.M. Royakkers (2011). *Ethics, technology, and engineering : an introduction*. English. United States: Wiley-Blackwell. ISBN: 978-1-4443-3095-3.

Poel, van de, I.R., J.N. Fahlquist, et al. (2012). "The problem of many hands : climate change as an example". English. In: *Science and Engineering Ethics* 18.1, pp. 49–67. ISSN: 1353-3452. DOI: 10.1007/s11948-011-9276-0.

Coeckelbergh, Mark (2020). "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability". In: *Science and Engineering Ethics* 26.4, pp. 2051–2068. DOI: 10.1007/s11948-019-00146-8.

Commission, European, Content Directorate-General for Communications Networks, and Technology (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office. DOI: doi/10.2759/002360.

Sio, Filippo Santoni de and Giulio Mecacci (2021). "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them". In: *Philosophy and Technology* 34.4, pp. 1057–1084. DOI: 10.1007/s13347-021-00450-x.

Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi (forthcoming). "Accountability in Artificial Intelligence: What It is and How It Works". In: *Ai and Society: Knowledge, Culture and Communication*.