# My Beautiful Paper About Stuff

Davide Rigamonti

June 2023

Fix title

Write abstract

## Todo list

## 1 Introduction

The historical model for determining the responsibility of a company or an individual that develops an AI system does not hold true 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 capable of autonomous action.

The main goal of this article is to underline the presence of a responsibility gap (Matthias 2004) when analyzing all kinds of autonomous systems, particularly those capable of "learning". It aims 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 insights. However, many questions still remain unaswered to this day. 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 achieve this, section 2 will begin by giving semi-formal definitions of *responsibility* (including some peculiar variations) and discuss 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.

## 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. It is crucial to acknowledge the presence of derived terms that represent different "flavors" of *responsibility*; in this section, our focus will be on the specific expressions that are most relevant to this paper. Nevertheless, it is also 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 Communications Networks, and Technology 2020 as cited in Novelli, Taddeo, and Floridi forthcoming); it is essential to note the presence of an authority in charge of supervising the conduct of the agent held accountable and thus, closely aligning with 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. In addition, it is also possible to identify the *patient of responsibility* figure as someone that is affected by the actions of the agent and is entitled to demand accountability for those actions (Coeckelbergh 2020).

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 achieved, the second is applicable in the event that something undesirable has already happened, in case of which it 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; specifically, *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 is easy to see how these frameworks fit quite well with most engineering tasks due to the fundamental instrumentality of the produced artifacts. For example, in the field of weapon systems, we can see that if the computerized aiming aid of a mortar happened to point the armament towards the wrong direction, and this happened due to an error in the code dedicated to the calculation of coordinates, once the root cause is assessed, it wouldn't be unthinkable to ascribe the responsibility to the programmer that wrote the code (provided that there were no other failures within the system, that the error was imperceptible to a human supervisor and that the calculations performed by the system don't involve machine learning or similar techniques). This happens because the code that runs on the system is considered an artifact with a precise function to carry out; the programmer knows this function and writes code that specifically performs the given task. If the artifact presents incorrect behavior it is reasonable to assume that its designer should be held accountable for the possible damages, thus resulting in the fulfillment of the *faultiness* considering their expertise and potential to anticipate such misbehavior.

## 3 The Responsibility Gap

In his seminal work, Andreas Matthias discusses the exisence 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 and foremost, it is important to comprehend in which ways autonomous learning systems differ from traditional engineering artifacts. In most instances, a learning automaton is able to produce outputs following a certain learned logic that wasn't directly coded by a programmer but has been "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 preconditions for ascribing responsibility (and even more so accountability) in traditional frameworks. However, when facing a system that has the properties of a *black box*, we cannot assume that anyone would be able to know its inner workings, not even 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.

An interesting case of acknowledging the responsibility gap and creating moral grounds to face it can be seen in the domain of Lethal Autonomous Weapon Systems (LAWS). In particular, the issue arises and shows its criticality due to the fact that systems in this domain are specifically designed to take human lives in an autonomous manner. In an unfortunate scenario, it may be possible that innocent civilian lives could be lost due to a minor error. When a human is directly employing a weapon, it is safe to assume that they bear the moral responsibility for any supposed error (assuming the weapon doesn't present any malfunction). However, if the weapon is operating on its own, suddenly the lines blur and in the event of an unfortunate incident, the process of attributing responsibility is no longer clear. To mitigate this problem, the International Committee of the Red Cross (ICRC) states that *meaningful*, *effective* or *appropriate* human control over weapon systems is needed and should be maintained in order to properly ascribe responsibility (Red Cross 2018). Nevertheless, no proper definition for these terms is given and many scholars have focused on defining and operationalizing the notion of *meaningful human control* in recent years (Verdiesen, Sio, and Dignum 2020; Ekelhof 2019).

# 4 Societal and Technological Consequences

In the previous section we have observed that applying traditional responsibility frameworks to AI systems may be an unwise decision, we will now examine the possible consequences of this choice. With technologies that are constantly being improved upon and a society that seems to be lagging behind it is important to understand that some issues can't be left unattended. The array of possible negative outcomes that may result from a mishandling of the core problem is wide, as there are numerous different types of imbalances that may significantly impact the resulting consequences in various ways.

If we decide to neglect the responsibility gap, the first undesired outcome becomes apparent: companies would be incentivized to distance themselves from their responsibility roles, shifting them on end users instead, resulting in frequent overlap for the *patient* and *agent* figures. As much as this solution is dubious, it is not completely erroneous from a moral responsibility perspective (assuming that the system provider takes enough care in making sure that the end user has a certain level of knowledge about the system which is, in itself, another crucial issue) but its true criticality can be seen when observed from a blameworthiness perspective. In this scenario the underlying culpability gap (Sio and Mecacci 2021) is even more apparent: the end user realistically doesn't possess the same amount of knowledge about the system as the engineer who designed it, consequently the end user is unfairly being asked to take the blame for the system's actions.

On the other hand, if an increasing number of authorities opt for enforcing a strictly traditional conception of responsibility for autonomous systems, we may observe the opposite effect: companies could lose interest in developing and utilizing new AI technologies, considering them too hazardous or outright unprofitable given their inherent unpredictability. In this scenario, AI research would be confined to theoretical and controlled environments, thus limiting its real life applicability, impeding the development process and depriving humanity of the benefits that could have been derived by the advancements in this field. To date, a lot of autonomous systems have already been deployed and internalized by society as useful tools for the daily use; removing them would impact the livelihood of many people, thus generating an additional moral dilemma on top of the issues that have been mentioned previously.

It is possible to identify a set of consequences that is strictly correlated to the field of LAWS and explores the eventuality of restrictions in their use. Assuming a simplified and theoretical scenario, if an influential geopolitical authority makes the decision to outright ban the use of autonomous killing machines in military operations due to the moral and responsibility implications that their use entails (Sparrow 2007), there will not be any guarantee that individuals and governments that are less observant of those values will not employ those technologies. This issue is amplified from the fact that LAWS are fairly accessible and their moral implications are less obvious compared to other regulated weapons such as nuclear and biological armaments. To conclude, we can identify two possible unfavorable outcomes in this scenario: either the restrictions won't be considered valid in practice, or the entities that decide to uphold the moral regulations will be at a perpetual disadvantage.

# 5 General Approaches

This paper will try to present a comprehensive approach that highlights a range of general responsibility practices for learning autonomous systems. This isn't meant to be a flawless or groundbreaking solution to any of the problems that have been identified so far as our approach focuses on the analysis and comparison of preexisting ideas.

Analyzing the consequences found in the previous section through a sociotechnical lens is fundamental for fully understanding the whole extent of the matter at hand (Theodorou and Dignum 2020; Novelli, Taddeo, and Floridi forthcoming); realizing that companies, corporations and their profit-driven approach have become the main driving force of AI development in the past years is the starting point for making any meaningful change. For a sociotechnical approach, it is important to not only manufacture technical solutions but also take into account the underlying social structure, aiming towards a joint optimization of the two components. This represents the main pitfall of the *solutionist* approach, as it implies that a general and absolute solution to the problem can be achieved solely through the implementation of new legal or technical tools (Morozov 2013; Stilgoe 2018; Sio and Mecacci 2021).

In their comprehensive analysis, Sio and Mecacci give an in-depth description of the main approaches to the central problem; together they identify the two main perspectives located at the two opposite extremes of the debate: *fatalism* and *deflationism*. While fatalism looks at the responsibility gap as an unsolvable dilemma, seeing the development of learning systems and human moral responsibility in strict mutual exclusion (Matthias 2004); deflationism takes the opposite stance and, while admitting that responsibility gaps exist, tries to portray their presence as an inevitable side effect that could be accepted and internalized by society since (as it happened with many technologies in the past) the benefits outweigh the drawbacks (Hayenhjelm and Wolff

2012; Simpson and Müller 2015).

Another compelling viewpoint is the *quasi-responsibility* approach (Stahl 2006), which explores the idea of considering autonomous systems as pseudo-agents of responsibility; this solution distinguishes itself from other approaches due to its radical nature, capable of solving the issue at its core. However, reach this conclusion is only possible under the assumption that the solution does, in fact, exist; assigning artificial moral values to a machine is not an easy task for various reasons and even then, from a human perspective it is not possible to hold an artifact blameworthy in the same way as a human might be, thus leaving a hypothetical victim of the system's actions without proper explanation and compensation.

## 6 Contrasting Viewpoints

Through a comprehensive analysis of contrasting viewpoints we can shed light on the issues of our thesis and hopefully reach solid conclusions.

One of the earlier formulations of *deflationism* (Hayenhjelm and Wolff 2012) that was analyzed in the previous section clearly goes against our main thesis as it refuses to acknowledge the responsibility gap as meaningful. However, the deflationist ideology evolved over time to accept the notion of responsibility gap and focus on other aspects of its original theory (Simpson and Müller 2015); these result grants some empirical credibility to the foundation of our thesis.

A possible counterargument to the main thesis is that trying to construct a new framework with new rules is superfluous as the classical one can be adapted to accommodate autonomous systems with the help of the inherent flexibility of law when interpreted by a court. The previous claim makes some strong assumptions on the quality, effectiveness and homogeneity of law systems around the world and across different cultures. However, our main thesis still holds since we would like to approach the issue on an ethical level of thought, where legal liability is just a part of the bigger picture.

Another important point to delve into is the consideration that there exists a variety of explainability tools capable of providing insight into the AI decision-making process. One example of such a tool is Grad-CAM (Selvaraju et al. 2016), which is specifically designed for deep neural networks dedicated to image processing applications. While it would be fair to say that autonomous systems are not entirely black boxes given the existence of these tools that can partially demystify the inference process of some AI models, there are still some major flaws to this argument. Firstly, analyzing the decision-making process of an autonomous system on a case-by-case scenario is cumbersome and also impossible if the system has to take split-second decision (as it may be the case with a LAWS system), thus relegating their use only to support passive responsibility audits. Even if we wanted to use explainability tools to assist the designers of an autonomous system to perfect its decisions, we could never percieve the system as an absolute white box due to the fact that the primary appeal of a learning autonomous system is strictly correlated to its inherent opacity. This opacity arises from the fact that want to solve tasks that are either too complex or too difficult to express in a formal way. Furthermore, having full knowledge regarding the decision making process of an AI system would imply that we could completely replicate it by using a tree of if-else statements, which is a deterministic process, thus reliable; obviously, this is under the assumption that the training and inference steps of the system are separate (not applicable to scenarios similar to reinforcement learning).

## 7 Conclusions

In this paper we discussed the various intricacies of the term responsibility and its alternative formulations, while mentioning some of the traditional ethical frameworks to manage responsibility and their philosophical grounds. We introduced the notion of *responsibility gap* and we observed the fact that classical engineering frameworks inevitably fall short when dealing with autonomous systems, which can be considered the core point of this paper. Once we recognized the existence of the responsibility gap, we turned our attention to potential solutions, analyzing and comparing various ideas present in the literature. Lastly, we took into account some counter arguments that challenged the feasibility of sociotechnical solutions and questioned the actual existence of a responsibility gap.

In conclusion, addressing the responsibility gap is not an easy task, and acknowledging its presence in modern society is the first step towards the creation of a new responsibility ecosystem for autonomous systems. There needs to be a collective effort to face the problem in the correct way while it is still possible to make changes to regulations and standards without causing destabilization in society. Autonomous systems are becoming more widespread and pervasive with each day that passes, and soon enough there might come a time where we will need to come to terms with the choices made in the past, at that point the presence of a solid foundation for responsibility management in learning systems will be of utmost importance.

# 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.

Stahl, Bernd Carsten (2006). "Responsible Computers? A Case for Ascribing Quasi-Responsibility to Computers Independent of Personhood or Agency". In: *Ethics and Information Technology* 8.4, pp. 205–213. DOI: 10.1007/s10676-006-9112-4.

Sparrow, Robert (2007). "Killer Robots". In: *Journal of Applied Philosophy* 24.1, pp. 62–77. DOI: 10.1111/j.1468-5930.2007.00346.x.

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.

Hayenhjelm, Madeleine and Jonathan Wolff (2012). "The Moral Problem of Risk Impositions: A Survey of the Literature". In: *European Journal of Philosophy* 20.S1, E26–E51. DOI: https://doi.org/10.1111/j.1468-0378.2011.00482.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0378.2011.00482.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0378.2011.00482.x.

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.

Morozov, E. (2013). *To Save Everything, Click Here: The Folly of Technological Solutionism*. PublicAffairs. ISBN: 9781610391399. URL: https://books.google.it/books?id=fdggBahA1qsC.

Simpson, Thomas W. and Vincent C. Müller (Aug. 2015). "Just War and Robots' Killings". In: *The Philosophical Quarterly* 66.263, pp. 302–322. ISSN: 0031-8094. DOI: 10.1093/pq/pqv075. eprint: https://academic.oup.com/pq/article-pdf/66/263/302/7441189/pqv075.pdf. URL: https://doi.org/10.1093/pq/pqv075.

Selvaraju, Ramprasaath R. et al. (2016). "Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization". In: *CoRR* abs/1610.02391.

arXiv: 1610.02391. URL: http://arxiv.org/abs/1610.02391.

Red Cross, International Committee of the (Apr. 2018). "Ethics and autonomous weapon systems: An ethical basis for human control?" In: URL: https://www.icrc.org/en/download/file/69961/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf.

Stilgoe, Jack (2018). "Machine learning, social learning and the governance of self-driving cars". In: *Social Studies of Science* 48.1. PMID: 29160165, pp. 25–56. DOI: 10.1177/0306312717741687. eprint: https://doi.org/10.1177/0306312717741687. URL: https://doi.org/10.1177/0306312717741687.

Ekelhof, Merel (2019). "Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation". In: *Global Policy* 10.3, pp. 343–348. DOI: https://doi.org/10.1111/1758-5899.12665. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1758-5899.12665. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12665.

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.

Theodorou, Andreas and Virginia Dignum (Jan. 2020). "Towards ethical and socio-legal governance in AI". In: *Nature Machine Intelligence* 2, pp. 1–3. DOI: 10.1038/s42256-019-0136-y.

Verdiesen, Ilse, Filippo Santoni de Sio, and Virginia Dignum (2020). "Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight". In: *Minds and Machines* 31.1, pp. 137–163. DOI: 10.1007/s11023-020-09532-9.

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*.