AI and Ethics (2021) 1:113–117
https://doi.org/10.1007/s43681-020-00009-0

OPINION PAPER

Responsible AI and moral responsibility: a common appreciation

Daniel W. Tigard[1]

Received: 19 August 2020 / Accepted: 2 September 2020 / Published online: 6 October 2020

Abstract
Responsibility is among the most widespread buzzwords in the ethics of artificial intelligence (AI) and robotics. Yet, the term often remains unsubstantiated when employed in these important technological domains. Indeed, notions like 'responsible AI' and 'responsible robotics' may sound appealing, for they seem to convey a sense of moral goodness or ethical approval, thereby inciting psychological connections to self-regulation, social acceptance, or political correctness. For AI and ethics to come together in truly harmonious ways, we will need to work toward establishing a common appreciation. In this commentary, I breakdown three varieties of the term and invoke insights from the analytic ethics literature as a means of offering a robust understanding of moral responsibility in emerging technology. While I do not wish to accuse any parties of incorrect usage, my hope is that together researchers in AI and ethics can be better positioned to appreciate and to develop notions of responsibility for technological domains.

Keywords Responsible AI · Responsible robotics · Technology ethics · AI ethics · Moral responsibility

1 Introduction

'Responsible AI', 'responsible robotics', 'responsible research and innovation', 'responsible technology': these notions have garnered widespread attention in recent years, within and beyond academic settings [3, 9, 16, 17, 19]. To a large extent, the growing popularity of the buzzwords is understandable. A great deal of uncertainty, and perhaps anxiety, along with efforts to quell the fears has arisen in discussions of emerging technologies, particularly surrounding AI and robotics. Yet, the idea of responsibility is often unsubstantiated in these discussions [6], and indeed, it appears to be employed as a placeholder for notions like moral goodness or ethical approval, thereby inciting psychological connections to self-regulation, social acceptance, or political correctness.

Being responsible is certainly much more than being morally good, and responsibility may well be ascribed to things which are far from being ethically approvable. To be sure, an individual might be appropriately considered responsible for committing moral atrocities. Accordingly, for AI research and ethics to come together in truly harmonious ways, researchers across disciplines will need to work toward establishing a common appreciation of this key concept. In this commentary, I breakdown three varieties of responsibility and invoke insights from the analytic ethics literature as a means of offering a robust understanding of moral responsibility for applications to technology. With this agenda, my goal is not to accuse anyone of incorrect usage. Rather, I aim to help all parties—technology developers, policymakers, social scientists, and ethicists alike—to better appreciate notions of responsibility in technology. My hope is that this conversation will be continued as we further develop what it means for humans to be responsible in our design, development, and use of technology, and what it might mean for technology itself to be held responsible.[1]

2 Varieties of responsibility

In everyday language, we tend to say things like 'Alex is a responsible young adult.' No doubt, this phrase seems to convey something positive or desirable about the subject. But what happens when Alex does something clearly

✉ Daniel W. Tigard
daniel.tigard@tum.de

[1] Institute for History and Ethics of Medicine, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany

[1] Note that I have used two subtly but importantly different locutions: *being* responsible and being *held* responsible. I discuss this distinction and its application to technology at greater length in Tigard [14].

careless—or irresponsible, as it were—say, driving drunk? A hopeful, forgiving parent will still be inclined to think 'In general, Alex is a responsible young adult.' Imagine that Alex continues to drive drunk and on one terribly unfortunate day, this results in the death of an innocent pedestrian. Do we still say Alex is a responsible young adult? Probably not, considering the accident and repeatedly careless behavior. But is Alex responsible *for* the pedestrian's death? Putting aside questions of ultimate causes (for example, whether it was Alex or alcohol that killed the pedestrian), some who deem Alex irresponsible will still say Alex is responsible for the death, or that as a driver Alex *has a responsibility*. I believe it is precisely this variety of uses that underlies much of the confusion about—and longing for substance in—the terms 'responsible AI' and 'responsible robotics.' Thus, to get clearer on these popular slogans, it seems fruitful to breakdown the three varieties, then consider what we can learn from each.

First, just as we often speak of young adult humans as responsible, we see growing discussion of AI and robotics as potentially responsible, where 'responsible' just means something akin to *the tendency or hope of behaving in positive, desirable or socially acceptable ways*.[2] For brevity, I will refer to this as the normative interpretation since it clearly imparts a value, namely a mark of endorsement. Second, we commonly refer to individuals or collections (such as societies or governments) as having responsibility, which seems to indicate something like *having a duty or obligation*, perhaps one with social, political, or moral significance.[3] I will refer to this as the possessive interpretation since with it we are ascribing or even bestowing upon others the possession of an immaterial property or quality. Third, we have the idea that individuals—perhaps also groups and institutions—can be *worthy of our responses in light of some action, intention, or outcome, in virtue of some relevant capacity*.[4] I will refer to this as the descriptive interpretation since here we are describing in detail the potential status of the subject, namely as the source or cause of something that happened. Notice that we can still evaluate what happened and the extent to which the subject is (or should be) responsible, and in doing so, we appeal to norms and values. This may seem to cross into a more normative reading. But the point is that on the descriptive interpretation, we are using the term 'responsible' in a way that does not (yet) impart a judgment of whether or not responsibility itself is positive or negative, unlike the normative interpretation where it is plainly good to be responsible and bad to be irresponsible. What can we learn from each of these interpretations when applied to AI and robotic technologies?

3 Normative responsibility

I want to begin by suggesting, perhaps rather provocatively, that the normative interpretation of responsibility when applied to AI and robotics can be the most pernicious of the three varieties. That is, deeming AI and robotics responsible in an effort to convey moral goodness, social acceptance, and so on, carries the highest risk of harm compared to the possessive and descriptive interpretations. The reasons for this potentially harmful use can be seen both in theory and in practice.

Consider, as described above, that young adults who are deemed responsible can nonetheless behave in morally atrocious ways. A murderer, for example, may well have been quite reliable, even trustworthy. What was it, then, that earned them the mark of responsibility? As I suggested, on this interpretation, what we mean when we refer to persons as responsible is that they have a tendency to behave—that we hope and trust that they will behave—in positive, desirable or socially acceptable ways. Of course, this sort of ascription is built upon a vast generalization of past observations; but actualities are not always consistent with usual tendencies. Just as individual humans can deviate from the behaviors they are observed as generally engaging in, machines too can deviate from their usual behaviors [5]. In either case, if responsibility is simply an indication of tendencies toward desirable or acceptable behavior, it seems that our recourse in light of any harm is simply to think of the subject as *less* responsible. In other words, 'responsible AI' and 'responsible robotics' may sound appealing in theory, but the mantras alone teach us little about how to remedy the immediate and future effects of actual behaviors exhibited by technological devices, developers, or users.

Perhaps we would do better to understand what features, in practice, are generally included in the normative mark of responsibility for technological domains. Consider, for example, the Foundation for Responsible Robotics (FRR), a Dutch non-profit organization headed by technology ethicist Aimee van Wynsberghe.[5] The FRR has developed and aims to deploy a "Quality Mark for (AI Based) Robotics"—a label that can help consumers to support societal values, and producers to earn trust and presumably profit. As van

[2] Along with public-facing organizations like the "Foundation for Responsible Robotics" which I will discuss, this interpretation is found in works such as [9], [2], and [3].

[3] This interpretation can be seen, for example, in [4], [8], [18], [3], and [7].

[4] Here I follow the formula offered by Shoemaker (12, p.17): "To be a responsible agent is to be worthy of X for Y in virtue of Z."

[5] Details can be seen here: https://responsiblerobotics.org/quality-mark/. Accessed 11 Sept 2020.

Wynsberghe explains, the label would be applied much like the 'Fair Trade' certification, where companies can earn a stamp of approval based upon an assessment of compliance with several guiding values. In the case of the FRR, those values are security, safety, privacy, fairness, sustainability, accountability, and transparency. To be sure, many of these same values are already touted by major technology corporations at the forefront of AI development, such as Google and Microsoft.[6] And while I do not want to diminish the noble efforts of the FRR and any like-minded organizations, the concern, of course, is that when for-profit companies promote such values, it may be little—if anything—more than an advertising campaign.

As the public's awareness turns more and more toward the effects and prevalence of AI, those with an interest in developing AI-based products are scrambling to produce whitepapers and public-facing documents vowing to protect social goods.[7] And here is where we see the potential practical harms of normative labels like 'responsible AI' and 'responsible robotics'. What we do not want, I assume, is for organizations to succeed in "ethics washing"—promoting terms like 'responsibility' in the service of deploying softer self-regulatory mechanisms and thereby instrumentalizing societal values as a means to greater shares of the technology markets.[8] Responsibility should be more than a positive mark of endorsement.[9]

4 Possessive responsibility

A great deal of technology ethics literature shows that 'responsibility' might indicate something other than moral goodness or ethical approval. Often it is thought to be a property or quality that a person or group can possess—say, in light of their development or use of emerging technologies. For instance, consider a recent work by Kirsten Martin [7], which aims to identify "whether firms developing algorithms *have a responsibility* for algorithms when in use" (p.836, italics added). Similarly, Virginia Dignum begins her book *Responsible Artificial Intelligence* (2019) by considering a wide array of stakeholders—"researchers, developers, manufacturers, providers, policymakers, users"—and claiming that "We all have different responsibilities" (p.v).

On these and many related accounts, we see no direct normative evaluation with the appeal to responsibility. That is, unlike on the normative interpretation, *possessing* a responsibility toward AI or robotic technologies might be positive or negative, depending upon the extent to which one "lives up" to those responsibilities, so-to-speak. However, notice that there is at least an indirect suggestion that fulfilling one's responsibilities is good, and to not do so is bad. For example, if a developer has a responsibility to design an algorithm in a way that produces fair outcomes, where the algorithm ends up producing clear biases, we can say that person has failed to live up to the responsibility of fairness in algorithmic design. But here we see that responsibility resembles little—and perhaps nothing—beyond a duty or obligation. Indeed, David Gunkel [4], upon stating that the key question for responsible robotics is where to locate responsibility, says "Who or what, in other words, can or should assume the obligations—the burden or duty—of answering for what a robot does or does not do?" Similarly, Dignum [3, p.54] maintains that "responsibility is the duty to answer for one's actions" and it is allegedly possessed by the agent before an action is undertaken. As she explains, "When a person delegates some task to an agent, be it artificial or human, the result of that task is still the responsibility of the delegating person" [3]. Thus, when outsourcing tasks to AI and robotic technologies, we retain possession of the duty to provide answers.

What we learn on the possessive interpretation is that responsibility, for some, remains uniquely human. If having a responsibility is to have a duty or obligation—regardless of what those duties or obligations entail—then we cannot say AI systems or robots have responsibilities unless we are prepared to accept that AI or robots can have duties or obligations.[10] Yet, it seems that responsibility must be more still, something beyond both a positive mark of approval and the possession of duties or obligations. After all, it becomes quite trivial to say of the various stakeholders—AI developers, policymakers, and so on—that they all have different duties or obligations. What more can we learn about responsibility when applied to AI and robotic technologies?

5 Descriptive responsibility

In contemporary analytic ethics literature, responsibility is treated widely—if not unanimously—as a neutral term, and as a concept that involves much more than duties and obligations. When we posit that some agent is a possible locus of responsibility, we must be prepared to address a

[6] See, for example, Google's 'Responsible AI Practices': https://ai.google/responsibilities/responsible-ai-practices/. Accessed 11 Sept 2020.

[7] For a thorough collection of these sorts of documents, by corporate and governmental institutions, see Ryan and Stahl [11].

[8] See Bietti [1] for a helpful discussion of "ethics washing" and "ethics bashing".

[9] For example, along with their seal of approval, the FRR aims to monitor and make compliance recommendations.

[10] For insightful discussion, see Nyholm [10] who suggests that our use of robots might create "obligation gaps".

number of key questions. First, what is the agent in question allegedly responsible for? Above, I suggested that on this third, descriptive interpretation, responsibility is a matter of being worthy of others' responses in light of some action, intention, or outcome, in virtue of some relevant capacity. To clarify, we may well extend this list to include an individual's decisions, attitudes, or even "quality" of will, as it is often framed in the ethics literature.[11] Granted, the process of ascribing responsibility becomes quite complicated when considering these sorts of details, but responsibility is indeed a complex notion and its intricacies deserve attention if we are interested in meaningful applications to technologies like AI and robotics. At present, the point to be made is that, on the descriptive interpretation, where we are describing the status of someone (or something) as responsible, we get a sense of what that person (or thing) is responsible *for*. We will also want to get clearer on, at least, two more questions: In what ways are they responsible? And why?[12]

Consider again the example, above, where a developer allegedly has a responsibility to design an algorithm in a way that produces fair outcomes. On the possessive interpretation, it appeared safe to say that that designer has a duty or obligation. In fact, on the descriptive interpretation too, duties and obligations, along with standards and expectations, play a key role in ascribing responsibility. However, on the descriptive interpretation, we are pressed to investigate much more than what the individual in question possesses. For instance, we might say Developer X is responsible for ensuring fairness in algorithmic design, meaning X is being held to an expectation that the algorithm succeeds in producing fair outcomes. Further, addressing the ways in which that developer is responsible, we might maintain, for example, that where this expectation is not met, the individual will be worthy of our demand for answers.[13] Accordingly, holding an individual *answerable* for failing to meet the demand of algorithmic fairness is a specific way we can respond—and with this response, we hold the individual responsible in ways not seen by simply deeming them a responsible developer, or by saying they have a responsibility. Notice, also, that the ways in which we hold others responsible, on this account, can and should vary, depending upon the target of our responses and what exactly we are responding to.[14]

Finally, when ascribing responsibility either to developers, users, or perhaps even AI and robotic technologies themselves, we should have a clear sense of why that person or thing is worthy of the responses we are deploying. Here, again, is where the descriptive interpretation can help us to achieve a robust understanding of responsibility. Consider once more that when Developer X fails to meet the expectation of algorithmic fairness and we respond with a demand for answers, we do so because that developer is seen as best positioned to provide answers in virtue of their awareness of the development process. In other words, the ways we respond—such as demanding answers—are justified by the individual's relevant capacities. Appealing to capacities, in this way, helps to make sense of why developers might be held responsible in ways that users are not, why corporations (or their CEOs) might be held responsible—perhaps via financial penalties—in ways that individual developers are not, and so on. In short, what we learn on the descriptive interpretation is that responsibility can be a detailed process of addressing concrete actions and events, for distinct reasons, and engaging with a diverse array of stakeholders depending upon their specific role and capacities.

6 Conclusion

Responsibility for AI and robotics is not simply a matter of applying positive marks of approval, even where societal values appear to be upheld, nor is it just an immaterial property that we may or may not ascribe. Admittedly, in such complex domains as emerging technologies, it also cannot be only a description of an individual's status, however detailed that description turns out to be. Continually in contemporary technology ethics, we see various ways in which the notion of responsibility is deployed, namely in an effort to manage the increasing effects of novel technologies in our lives. What I hope to have conveyed with this brief commentary is not that anyone is necessarily deploying the concept incorrectly. In fact, it is reassuring to consider that researchers across disciplines—whether working on developments in AI or applied ethics, or both—are making use of such complex ideas as responsibility. However, I hope also that the complexities will be given due attention and that we will work increasingly together to make sense of responsibility in today's emerging technologies.

Acknowledgements The author would like to thank the Bavarian Research Institute for Digital Transformation for funding the project "Responsible Robotics: Tracing Ethical and Social Aspects of AI-Based Transformations in Healthcare Work and Knowledge Environments" at the Technical University of Munich.

Funding Open Access funding provided by Projekt DEAL.

[11] The "quality of will" literature on responsibility generally follows from the work of P.F. Strawson [13].

[12] Again, I follow the formula offered by Shoemaker [12], see note 4.

[13] Here I have in mind Shoemaker's notion of *answerability*, which I apply to technological entities in Tigard [15]. For a related approach, see Coeckelbergh [20].

[14] As a brief contrast, consider that where the demand for algorithmic fairness *is met* (or even exceeded by unexpectedly positive results), we would reasonably commend that developer—and perhaps *then* grant the positive mark of 'responsible' AI development.

Compliance with ethical standards

Conflict of interest The author declares no conflicts of interest.

References

1. Bietti, E.: From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 210–219 (2020)
2. Boden, M., Bryson, J., Caldwell, D., et al.: Principles of robotics: regulating robots in the real world. Connect. Sci. **29**(2), 124–129 (2017)
3. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer, New York (2019)
4. Gunkel, D.: Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf. Technol. 1–14 (2017)
5. Illankoon, P., Tretten, P., Kumar, U.: Modelling human cognition of abnormal machine behaviour. Human Intell. Syst. Integr. **1**(1), 3–26 (2019)
6. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)
7. Martin, K.: Ethical implications and accountability of algorithms. J. Bus. Ethics **160**(4), 835–850 (2019)
8. Mittelstadt, B.: 'The doctor will not see you now': the algorithmic displacement of virtuous medicine. In: Otto, P., Gräf, E. (eds.) 3TH1CS—The Reinvention of Ethics in the Digital Age. IRights Media, Berlin (2017)
9. Murphy, R., Woods, D.D.: Beyond Asimov: the three laws of responsible robotics. IEEE Intell. Syst. **24**(4), 14–20 (2009)
10. Nyholm, S.: Humans and Robots: Ethics, Agency, and Anthropomorphism. Rowman & Littlefield, Lanham (2020)
11. Ryan, M., & Stahl, B.C.: Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J. Inf. Commun. Ethics Soc. (2020)
12. Shoemaker, D.: Responsibility from the Margins. Oxford University Press, Oxford (2015)
13. Strawson, P.: Freedom and resentment. Proc. Br. Acad. **48**, 1–25 (1962)
14. Tigard, D.: Artificial moral responsibility: how we can and cannot hold machines responsible. Cambridge Quarterly of Healthcare Ethics (forthcoming)
15. Tigard, D.: There is no techno-responsibility gap. Philos. Technol. (2020). https://doi.org/10.1007/s13347-020-00414-7
16. Umbrello, S.: Imaginative value sensitive design: Using moral imagination theory to inform responsible technology design. Sci. Eng. Ethics **26**, 575–595 (2020)
17. Van Oudheusden, M.: Where are the politics in responsible innovation? European governance, technology assessments, and beyond. J. Responsib. Innov. **1**(1), 67–86 (2014)
18. van Wynsberghe, A., Donhauser, J.: The dawning of the ethics of environmental robots. Sci. Eng. Ethics **24**(6), 1777–1800 (2018)
19. Zhu, W.: 4 Steps to Developing Responsible AI. World Economic Forum (2019). https://www.weforum.org/agenda/2019/06/4-steps-to-developing-responsible-ai/. Accessed 11 Sept 2020
20. Coeckelbergh, M.: Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci. Eng. Ethics **26**(4), 2051–2068 (2020)