Cambridge Quarterly of Healthcare Ethics (2021), 30: 3, 455–458
doi:10.1017/S0963180120001000

CAMBRIDGE UNIVERSITY PRESS

RESPONSE

Artificial Agents in Natural Moral Communities: A Brief Clarification

Daniel W. Tigard*

Institute for History and Ethics of Medicine, Technical University of Munich, 81675 Munich, Germany
*Corresponding author: Email. daniel.tigard@tum.de

Abstract

What exactly is it that makes one morally responsible? Is it a set of facts which can be objectively discerned, or is it something more subjective, a reaction to the agent or context-sensitive interaction? This debate gets raised anew when we encounter newfound examples of potentially marginal agency. Accordingly, the emergence of artificial intelligence (AI) and the idea of "novel beings" represent exciting opportunities to revisit inquiries into the nature of moral responsibility. This paper expands upon my article "Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible" and clarifies my reliance upon two competing views of responsibility. Although AI and novel beings are not close enough to us in kind to be considered candidates for the *same* sorts of responsibility we ascribe to our fellow human beings, contemporary theories show us the priority and adaptability of our moral attitudes and practices. This allows us to take seriously the social ontology of relationships that tie us together. In other words, moral responsibility is to be found primarily in the natural moral community, even if we admit that those communities now contain artificial agents.

Keywords: moral responsibility; moral agency; blame; machine ethics; artificial intelligence; human–robot interaction

What exactly is it that makes a person—namely, a fully functional adult human being—morally responsible? Is it a set of facts which can be objectively discerned, perhaps by looking carefully enough at the person in question? Is it something more subjective, a reaction to the person or context-sensitive interaction between her and others? This debate is not new, but it often gets raised anew when we encounter persons or things other than fully functional adult human beings. Considering the emergence of artificial intelligence (AI) and the idea of "novel beings" is, then, an apt invitation to revisit such inquiries into the nature of moral responsibility.

With my article "Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible,"[1] I aim to align myself with a strand of contemporary ethics that follows from the work of P.F. Strawson. In particular, I follow in the footsteps of David Shoemaker, whose 2015 book *Responsibility from the Margins* took as its starting point the "fact of our ambivalence" in how we respond to diverse individuals, such as psychopaths, persons with intellectual disabilities, and those on the autism spectrum, among others.[2] Rather than looking straight into a supposedly clear case of moral responsibility in an effort to discover its nature, Shoemaker helps us to see that we learn a great deal when we examine the marginal cases. In this way, an inquiry into how some AI systems *might* be considered morally responsible, I take it, represents an extremely marginal case—one that we can still learn from, nonetheless.[3]

I realize that the project of seeking moral responsibility in AI itself is largely an uphill battle. Our technological creations are simply not close enough to us in kind to be seriously considered potential candidates for the *same* sorts of responsibility we ascribe to our fellow human beings. With that in mind,

I was pleased to see a response offered by Marc Champagne,[4] but not surprised at the challenges it tries to put forward. In this brief paper, I want to clarify a few key points from my initial argument, particularly my reliance upon two competing views of responsibility. It may be, as Champagne suggests, that addressing the ontology of individual agency is necessary. Still, I stand with contemporary responsibility theorists who prioritize our moral attitudes and practices, and thereby take seriously the social ontology of relationships that inevitably tie us together. In short, moral responsibility is to be found primarily in the natural moral community, even, I suggest, if we admit that those communities now contain artificial agents.

To begin, I will clarify what I take to be the more objective view of responsibility, since, despite his support for the view, an extreme reading of it was expressed as a point of confusion in Champagne's paper.[5] So, what *does* it mean to say responsibility is an objective property? As I initially described it, this is the idea that *being* responsible is prior to being *held* responsible. Shoemaker calls this position the "B-tradition." Here, there is supposed to be some property (or set of properties) that *by itself* determines whether or not someone is responsible. One question, then, is just: What are those properties? I noted that knowledge and free will, or perhaps ill will, are very common candidates. If one knows he is doing wrong and does so freely, or with an ill will, he is morally responsible for it. Accordingly, if we say he is not morally responsible, it must be that he did not know or somehow was not free, or did not possess ill will after all.

Another question to be asked is: What role, if any, do our natural responses play? Proponents of the B-tradition can grant that our responses are important, but only as mechanisms by which we detect some underlying fact about the person's prior responsibility status. In this way, resentment and guilt, for example, are "epistemic markers"—they do not *constitute* the facts of responsibility.[6] Those facts or properties must be independent from our attitudes and practices. Indeed, this view is usually what Champagne appears to be pointing to when he claims that the "ontological issue is crucial" or "mandatory."[7] Undoubtedly, for those who consistently maintain the B-tradition, it will be quite difficult to really make sense of moral responsibility in artificial agents.[8] Why? Responsibility, on this view, looks to be inextricably intertwined with features we ordinarily believe to be unique to human agency—again: knowledge, free will, perhaps consciousness, empathy, and so on. Thus, if the ability to identify responsibility is morally required—say, for harms in warfare or medicine—it will appear that deploying AI in such domains is morally impermissible.

However, I do not think that is all that can be said about locating moral responsibility in artificial agents. In particular, responsibility can be seen as a process, or perhaps better, a social practice. The idea here is that *holding* responsible is prior to *being* responsible—the "H-tradition" or response-dependent view, in Shoemaker's terms.[9] On this picture, the key to responsibility is our natural responses; it is the fact that we hold others (and ourselves) responsible. And here, Champagne is right to insist that "we expect our best 'holding' practices to track real features, as opposed to being purely spontaneous ascriptions." Fortunately, our responses are not purely spontaneous. They are prompted by distinct events and circumstances, and by a great variety of agential characteristics. The proponent of the H-tradition can even readily acknowledge that our responses track *real* features of the target agent. The question is, then: Do those features *alone* constitute moral responsibility?

For those following the B-tradition, it is clear that nothing more is needed other than some relevant facts about the agent in question—that he knowingly did wrong, did so freely, possessed ill will, and so on. But notice what this view does to such morally loaded notions. It starts to look like wrongness, freedom, and ill will are as observable as purely descriptive features of the agent, like the fact that he has two arms and that he indeed knocked over an elderly person. As Neal Tognazzini aptly states, it is "not as if we can just make a list—Well, this guy was mean, he had no good reason to be mean, he knew what he was doing."[10] And even if we could make such a list, we would still need to know what it is about those features that enables us to appropriately hold others responsible, in particular, via blame.

Champagne posits that "When asked why we hold so-and-so responsible, we tend to answer without missing a beat that it is because so-and-so *is* responsible." This answer sounds peculiar to me, in part because I do not share the intuition that we would tend to say that, but also because it does not really answer the question. Why do I hold a good friend responsible for blowing off the plans we made? Because

it hurt my feelings and I expect better treatment. Perhaps also because calling him out on it will let him know that it hurt my feelings and that I expect better. It may also, hopefully, discourage this sort of thing from happening again. No doubt, if we look carefully enough at my inconsiderate friend, the facts we discover may well include things like he knowingly blew off our plans, did so freely and with an ill will. But, importantly, these facts alone do not fully explain why I hold him responsible. Imagine, for example, I no longer expect any better from this person, or for some reason never did. Indeed, to suppose that the facts of the target agent alone are enough to explain moral responsibility is to ignore the social ontology of the situation, namely the moral community in which we both participate. In this way, prioritizing our social practices and interactions certainly does not "dodge ontology"—quite the contrary. Proponents of the response-dependent view are able to explain key features of individuals and of our relationships, precisely, in Strawson's words, "by attending to that complicated web of attitudes and feelings which form an essential part of the moral life as we know it."[11]

I return now to the task of applying the response-dependent view to artificial agents. Consider a real-life example. In July of 2016, a shopping mall security robot—known as Knightscope K5—struck and ran over a 16-month-old toddler. The boy suffered no long-term injuries and K5's overseeing company was quick to issue a reassurance of their commitment to safety. Nonetheless, the boy's parents were understandably upset. In an interview, the mother expressed that "the robot did not stop at all," as if she had expected the 300-pound machine to stop upon hearing her screaming.[12] And naturally, such expectations might seem peculiar or forlorn—but it is worth noting, again, that some technologies are increasingly able to recognize and respond to our moral attitudes.[13] Does this sort of responsiveness qualify some devices as candidates for moral responsibility? Again, it would still seem peculiar. But unlike Champagne, I do not find it "unlikely that human culture will adapt."

In fact, whether we like it or not, it appears that we are already adapting to the inclusion of artificial agents in our moral communities, even if they are far from full members. Consider here the research suggesting that people cannot help but respond emotionally to humanoid robots, or the stories of soldiers truly bonding with military robots.[14] With these true-to-life accounts of the adaptability of human culture in mind, the overarching agenda in my initial paper was not necessarily to show that we *should* hold machines responsible. To be sure, I think there are often good reasons *not* to, namely when there are identifiable human associates—designers or users, for instance—who should "take responsibility," a notion that both Champagne and I have previously supported.[15] However, one of the key ethical problems of emerging technologies is that we may soon, if not already, face situations where there simply are no identifiable human associates, despite the occurrence of serious harm. This problem was the motivation for framing my account against the backdrop of the technological "responsibility gap." My main line of inquiry remains: What *can* we do about it?

To some extent, I agree with Champagne—namely that we often *want* "real moral responsibility." That is, we want someone to be *there*, but not just to possess a set of mandatory properties, rather, to hear and receive our demand for moral concern. We want there to be someone with whom we can relate, as a fellow member of the moral community. By analogy, Champagne says "In our hospitals, we do not want artificial nurses. We want real nurses." Yet, if I am suffering unnecessarily and my only available remedy is someone working and caring *as a* nurse, I will be pleased at the possibility of relief of my suffering, whether or not the imposter possesses official nursing credentials. Likewise, when I experience resentment or indignation at some perceived harm, ideally, I will find that the target of my attitude is capable of understanding and responding accordingly. But like hospitals, our moral lives very often present nonideal circumstances, and we would do well to remain open to new ways of promoting our wellbeing.

As things stand, it appears that we will increasingly encounter novel beings, AI and robotic systems which will surely not be capable of fully understanding our moral attitudes and practices. In this way, I do not claim that we can hold machines responsible in the same ways we hold fully functional adult human beings responsible. Still, it might be that some technologies are becoming capable of responding to us in ways that satisfy our natural propensity to engage in moral interactions. For this reason, I strongly suggest that we not push "outliers to the side"—for it is precisely these individuals and these relationships from which we can learn.

Notes

1. Tigard D. Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics*; available at https://doi.org/10.1017/S0963180120000985. forthcoming.
2. Shoemaker D. *Responsibility from the Margins*. New York: Oxford University Press; 2015.
3. That being said, I want to reiterate—as I pointed out in the initial paper—that with my inquiry into responsibility for *artificial* intelligence, I deviate from Shoemaker's investigation of *natural* subjects. Accordingly, I take full responsibility for any unbecoming distortions of his theory.
4. Champagne M. The mandatory ontology of robot responsibility. *Cambridge Quarterly of Healthcare Ethics*; available at https://doi.org/10.1017/S0963180120000997. forthcoming.
5. Note that using the notion of "extremes" to depict the two views is simply an analytic tool, a way of drawing definite distinctions. I do not believe many theorists hold one of these extremes; instead, seeing the two views along a continuum, or maintaining some combination, seems more plausible. In any case, it is unclear how my framing of the contrast "misconstrues the relation" as Champagne writes.
6. For a fuller explanation, see note 2, Shoemaker 2015, at 19–20.
7. Yet, on occasion, he says things like "once the jury has *found* one guilty, one *is* (and thus *was*) guilty" (emphasis in original), indicating support for a more constructivist reading, which helps my case for locating responsibility in our practices.
8. Hence my subtitle: How we can *and cannot* hold machines responsible. See note 1.
9. Shoemaker D. Response-dependent responsibility; or a funny thing happened on the way to blame. *Philosophical Review* 2017;**126**:481–527.
10. See Tognazzini's defense of contemporary Strawsonians, in Tognazzini N. Blameworthiness and the affective account of blame. *Philosophia* 2013;**41**:1299–312.
11. Strawson PF. Freedom and resentment. *Proceedings of the British Academy* 1962;**48**:1–25. Considering the enormous impact of Strawson's work, we see that Champagne is simply mistaken to think the objective view of responsibility has "been around for too long" to be dislodged.
12. Favro M. Mother says 16-month old son injured by security robot at Stanford shopping center. *NBC Los Angeles* 12 July 2016; available at https://www.nbclosangeles.com/news/national-international/15-Month-Old-Boy-Injured-By-Robot-at-Stanford-Shopping-Center-386544141.html (last accessed 13 Dec 2019).
13. See, for example, Ren F. Affective information processing and recognizing human emotion. *Electronic Notes in Theoretical Computer Science* 2009;**225**:39–50. Also, Knight W. Amazon working on making Alexa recognize your emotions. *MIT Technology Review* 2016.
14. See, for example, Parthemore J, Whitby B. Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness* 2014;**6**:141–61; also, Brezeal C, Scassellati B. How to build robots that make friends and influence people. *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 17–21 Oct1999, Kyongju, South Korea; Garreau J. Bots on the ground: In the field of battle (or even above it), robots are a soldier's best friend. *Washington Post* 6 May 2007.
15. Champagne M, Tonkens R. Bridging the responsibility gap in automated warfare. *Philosophy and Technology* 2015;**28**:125–37; Tigard D. Taking the blame: appropriate responses to medical error. *Journal of Medical Ethics* 2019;**45**:101–5. Tigard D. Taking one for the team: A reiteration on the role of self-blame after medical error. *Journal of Medical Ethics* 2020;**46**:342–4.

Cite this article: Tigard, D. W. 2021. Artificial Agents in Natural Moral Communities: A Brief Clarification. *Cambridge Quarterly of Healthcare Ethics* 30: 455–458, doi:10.1017/S0963180120001000