Cambridge Quarterly of Healthcare Ethics (2021), 30: 3, 448–454
doi:10.1017/S0963180120000997

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

The Mandatory Ontology of Robot Responsibility

Marc Champagne

Department of Philosophy, Kwantlen Polytechnic University, Surrey, BC, V3W 2M8, Canada
Email: marc.champagne@kpu.ca

We humans are adept at picking out members of our own species, but overly sensitive aspects of this "cultural intelligence"[1] have unforeseen consequences. Paint an angry face on a missile, for example, and it seems to not just track its target but "hunt" it.[2] This can lead to confusion about who to praise or blame when machines enhance or hinder our lives. Robert Williams, for instance, may have been the first person to die as a result of a robot,[3] but he was certainly not the first to die as a result of a machine—think, for example, of the dangerous working conditions of the early industrial revolution. Yet, is it more reasonable to blame a machine because it has an articulate "arm" as opposed to, say, a cast iron crane?

Robots are increasingly designed to look, sound, and/or react like us precisely because this similarity stirs our sentiments. According to Selmer Bringsjord:

> [W]e're headed toward realizing *Blade Runner*, a classic sci fi [*sic*] movie in which only an elaborate pupil-scanner (which detects the usual physiological correlate to an emotional response to provocative questions) enables one to distinguish androids from humans. […] [W]hat might be called the 'Double Blur' is […] the gradual blurring of the traditional boundaries between human persons and robots (or androids).[4]

Ethically, one might hold that, as the difference between robots and us diminishes, the warrant for treating robots like us increases. The problem, however, is that many people resist this line of reasoning. Even today, "[p]hysicians' dependency on helpware leads to questions of liability, of who exactly is to blame if a doctor accepts the wrong recommendation of an AI," and "[t]he issue of how an AI might be 'credentialled' is a complex one" that has "given rise to lawsuits."[5] Hence, Bringsjord surmises that "though this blurring will happen, significant differences between robots and human persons will persist."[6]

Do we suddenly become justified in treating robots like humans by positing new notions like "*artificial* moral agency"[7] and "*artificial* moral responsibility"[8]? I will answer no. Or, to be more precise, I will argue that such notions may become philosophically acceptable only after crucial metaphysical issues have been addressed.

Treating a moving chunk of matter "as if" it had intentions and beliefs can certainly be fruitful, from a predictive standpoint.[9] It can also be useful from a social standpoint. Nevertheless, such a stance mostly[10] dodges the ontological questions it raises. As Jerry Fodor cynically but rightly remarked, "[t]he great virtue" of justifying mental states based on their usefulness "is that you get all the goodness and suffer none of the pain: you get to use propositional-attitude psychology to make behavioral predictions […] but you don't have to answer hard questions about what the attitudes *are*."[11] Alas, those hard questions will not go away.

Normally, we don't dodge ontology; we tackle it head-on. 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. The target of correct moral evaluations is usually supposed to have some feature(s), independently of what

anyone thinks. If, for instance, you killed someone, this fact holds irrespective of whether anyone discovers it or appeals to your action to justify some subsequent punishment. You become a murderer the moment you murder someone, not the moment someone else calls you a "murderer." Such mind-independence, of course, is the hallmark of the real. Hence, we usually want those that we deem wrong to be wrong—in the demanding ontological sense of the verb "to be."

To gauge the importance of this ontological requirement, consider the following news story:

> Health officials in Quebec have discovered that a woman who had been working as a nurse and caring for hospital patients for 20 years was an impostor. […] Her ploy was discovered a few weeks ago when she enrolled in a training course. An official noticed that the age listed on her license number did not match up with her actual age. She was immediately suspended pending an investigation, which led to her dismissal. Before being exposed, the woman had worked in several departments of the hospital, including the operating room. A spokeswoman for the health authority said a search turned up no evidence the woman had a nursing degree.[12]

Upon learning of this story, someone could in principle reply: "But she performed her functions well, so the recent revelation of her lack of credentials makes no practical difference." However, the question of whether she is a "real" nurse—even though she walked like one, talked like one, looked like one, and so on—seems paramount. Indeed, even if there is no evidence anyone can point to in order to establish that the "as if" nurse was incompetent, people turn to a weaker modal idiom to ontologically ground their sense that she was not a real nurse. For instance, "Luc Mathieu, president of Quebec's Order of Nurses, said such cases of impostors are rare but troubling. 'It's very serious, because that person *could* have committed acts that *could* have had serious consequences for patients,' he said" (emphasis added).[13] This shows that the question of whether agents are *really* responsible is foremost on most people's minds when they ponder applied questions pertaining to morality. Such a concern with reality does not bode well for artificial versions of agency and responsibility.

Must agents really be responsible in order for us to blame or praise them? As Daniel Tigard points out, there are basically two ways of looking at the situation. On the one hand, there is the view according to which "moral responsibility is a process—we hold others responsible and, importantly, *holding* is conceptually prior to *being* responsible."[14] On this view, "the facts of responsibility are determined primarily by our responses, our attitudes and practices, and not the other way around. In other words, moral agency does not *really* matter for our determinations of responsibility."[15] This is the view at play in Isaac Asimov's short story about a boy on a lunar colony who loved his robotic canine more than real dogs.[16] On the other hand, there is a view that insists that "we must—and we often do—account for the agential status of the objects to which we respond. On this line of thought, responsibility is an objective property, and our determinations of it will be based largely upon our awareness of fundamental features of the object."[17] On this view, no amount of love by the boy or anyone else will alter the fact that the robot is a robot, not a dog.

Tigard portrays these two views as bookends on a spectrum and says "I want to avoid coming down definitively in favor of either extremity."[18] However, I believe this portrayal misconstrues the relation between "holding" and "being." The relation seems to be more like this: "I *hold* you to be R. My holding is correct. Therefore, you *are* R. Why are you R? Because *you* are R (not because *I* hold you to be R)." This line of reasoning is ampliative, not deductive. Even so, we expect our best "holding" practices to track real features, as opposed to being purely spontaneous ascriptions. So, it is not that objective responsibility takes things to an "extreme" (whatever that means). Rather, the objective view is typically regarded as the culmination of any successful ascription of blame/praise. Provisionally suspending judgment may help us impartially weigh the pros and cons of evidence. But, assuming a proper method of inquiry, once the jury has *found* one guilty, one *is* (and thus *was*) guilty. No human judgment can escape epistemic fallibility, but we have a crystal clear conception of the ideal alignment with the facts being sought.[19]

Being "a member of the moral community"[20] is thus necessary but not sufficient, since what Antoine de Saint-Exupéry said of love can be said of moral assessments, namely that they do not consist in looking

at each other, but rather looking in the same direction. In fact, I would argue that one becomes a genuine member of the moral community precisely once one commits to letting the facts, not one's peers, guide one's judgments. So, while it may be expedient to treat machines "as if" they were humans, the facts that constrain correct moral assessments cannot be altered by any social practice(s).

To see this, consider the following argument. Imagine that you are vacationing in China. Suddenly, to your surprise, you are detained by Chinese officials and charged with espionage. You did no such thing, so when they interrogate you regarding the whereabouts of some secret briefcase, you reply that you genuinely do not have the knowledge they seek. Of course, such a reply is indistinguishable from the one that would be offered by a spy withholding information. So, after the story of your capture and failure to cooperate is broadcast on state television, you find yourself blamed by over a billion people. Yet, despite this massive ascription of blame, you simply lack the mental state(s) that would make you really blameworthy. Now, like the foreigner accused of spying, robots lack the mental state(s) that would make them blameworthy,[21] even when those robots give replies indistinguishable from the ones offered by capable (conscious, adult, etc.) persons. If the social practices and attitudes surrounding a robot could suffice to render it "responsible" (responsible enough, at any rate, to merit the label or its cognates), then by analogy concerted blame by the Chinese would suffice to let you know the location of the missing briefcase. This is absurd. The conclusion of this argument, then, is that being *held* responsible is not *being* responsible.

Those who champion the notions of artificial moral agency[22] and moral responsibility[23] may not make much of this demand that the agency and responsibility at hand be real. I contend, however, that the ontological issue is crucial. Take, for a comparison, current debates about qualia. Qualia are, by definition, epiphenomenal. In other words, raw feels like tastes and colors do not contribute to the causal and inferential functions that we associate with cognition.[24] Now, Tigard writes that, "'if [an artificial moral agent's] architecture and mechanism allow it to do many of the same tasks' as human consciousness, it can be considered *functionally conscious*" (emphasis in original).[25] Yet, if analytic philosophers of mind—who, as a group, are hard-nosed naturalists—cannot rid themselves of the worry that perfect duplication of human behavior is not enough to conclusively track mentality, I doubt that ethicists will unproblematically accept notions of artificial moral agency/responsibility. Passing an "as if" Turing test did not suffice to settle descriptive concerns in philosophy of mind, so passing an "as if" test is unlikely to settle normative concerns.

The claim that "some who are punished *suffer the consequences*, even if they cannot experience the suffering of a natural moral agent" (emphasis in original)[26] involves different meanings of the word "suffering." Robots do not "suffer" in the experiential sense, so we will likely view robot trials in an even more absurd light than we currently see medieval European trials of inanimate objects.[27] Robert Sparrow thus has his finger on the pulse of humanity when he holds that, "[f]or punishment to be punishment […] its target 'must be capable of suffering.'"[28] That is why, despite uncoupling the causal and the moral in our "blank check" proposal, Ryan Tonkens and I require that, in the end, a human (who switched on a violent autonomous robot) suffer.[29] You have to have skin in the game. Robots don't. That may not look like much. But, if embodied perishability is the source of values, then not having anything vital to gain or lose from one's actions may prove decisive.[30]

Tigard asks: "[W]ho or what else, if anyone or anything other than fully functioning adult human beings, can truly possess moral agency? On the natural picture of moral agency, as I have in mind here, the short answer is no one and nothing."[31] This answer overlooks that, in almost all science fiction narratives, aliens are treated as moral agents. Consider that, while there had to be an episode of *Star Trek: The Next Generation* devoted to establishing whether Data (an android) is deserving of blame/praise,[32] the idea that Worf (an alien) is deserving of blame/praise literally went without saying. Our mythic projections about extraterrestrials thus reveal strong (and perhaps immutable) intuitions that privilege fleshy tissue.

That is not to say that we deploy moral vocabulary in a straightforward manner. The idea that "responsibility" has multiple meanings has been around since H. L. A. Hart, who distinguished between causal-responsibility, liability-responsibility, capacity-responsibility, and role-responsibility.[33] Likewise,

David Shoemaker's tripartite theory of attributability, answerability, and accountability[34] does justice to the fact that the game of praising and blaming is quite complex. Because some of the obstacles have to do with this complexity, it is tempting to think that technological advances will eventually let moral principles be applicable to artificial agents. For instance, we demand responsible agents to justify their actions. Independently converging with that demand, engineers and programmers now realize that a robot must explain *why* it is engaging in a task, since only in light of a goal can it (and others) evaluate whether the task was well or poorly carried out. To give an example, "sometimes workers not only have to pick the object, but also read a barcode. This implies orienting the object in a way that the reader recognizes such a barcode. In this case, speed cannot be the only measure of quality for the grasp,"[35] since a palm or finger placement occluding the crucial information would render the picked object essentially useless.

Admittedly, it would take relatively minor enhancements in data monitoring, storage, and retrieval for robots to convey in a more detailed way the results that they are trying to achieve. Since we tend to do this too, this ability would add another layer of mimicry. So, predictably, recent studies suggest that humans are more likely to trust a robot when, in addition to doing things correctly, it can explain what it is doing.[36] While the answers given by robots might tighten the link of causal-responsibility, we should not conflate such differential responses with the responsiveness to reasons distinctive of capacity-responsibility. Robots may reliably respond to situational inputs by outputting actions that humans deem good, but then again "[a] chunk of iron reliably responds to some environments" by rusting, even though "[t]he chunk of iron is not conceiving its world as wet when it responds by rusting."[37] Overlaying that ability with narration does not magically turn the efficient causation into teleology. Robots may be sophisticated enough to report "what" goal they are pursuing, but this does not mean that they know "why" they are pursuing it. After all, "[t]he obedient and dutiful child is not yet responsible, at least not until she starts to exercise her own judgment," and "the well-trained animal never will be responsible."[38] We can add the complicated robot to that list.

More problematically, it is entirely possible for a robot to perform tasks using one set of guidelines/principles yet issue verbal reports that refer to another set of guidelines/principles. As roboticists aver, "model components that foster the most trust do not necessarily correspond to those components contributing to the best task performance. This divergence is possible because there is no requirement that components responsible for generating better explanations are the same components contributing to task performance; they are optimizing different goals."[39] Robots can simply be telling us what we want to hear, so it would be an indulgence in delusion to think that their reports conclusively track anything resembling "motives."

The replicants from the *Blade Runner* movies are designed to supply reasons for their actions when prompted. In fact, they replicate us so well that pupil-scanners are required to detect their presence. Now, reacting differently to a pupil-scanner is a minute difference, so maybe in time we will tire of tracking that difference and simply decide that treating replicants like regular humans is more economical.[40] Yet, there are also reasons to think that the ontological difference between humans and robots will be rendered even more salient. In our living rooms, we don't want artificial wood. We want real wood. In our hospitals, we don't want artificial nurses. We want real nurses. Even in philosophical debates far removed from practical consequences—and professionally aware of how slippery the word "is" is—it bothers us whether we are interacting with artificial minds or real minds. By induction, humans may continue to demand real moral responsibility.

The notion of artificial moral agency is premised on the idea that humans have the power to devise, adopt, and find convincing new moral concepts to meet new moral conundrums. Yet, outliers to the side, it is unlikely that human culture will adapt in a way that makes room for artificial moral responsibility. Historically—and thus psychologically—the attitude that focuses on practical differences (or lack thereof) instead of essential natures is a fresh coat of paint on a rock that is a mile deep.[41] As a result, we cannot ignore the human appetite for a more ontologically committed explanation of what it means to deserve blame or praise. For instance, even if an AI can eventually do a better job of being the leader of a country, it is unlikely (and undesirable?) that an AI will one day be elected. *Flesh matters, even if we do*

not (yet) fully know why it matters. That is why, "[f]or some, it seems that the prospect of losing our grip on responsibility, even for fewer harms, is worse than knowing exactly who is responsible" when things take a turn for the worse.[42] Indeed, if there is one result that investigations into applied ethics have established, it is that consequentialism moves only a (pathological?) few.[43]

Tigard tries to "move beyond natural conceptions of moral responsibility by showing that, often, our attitudes and practices are adaptable to AI systems."[44] Our attitudes and practices are adaptable to a certain extent, but the "more objective view, wherein responsibility is a property rather than a process"[45] has simply been around for too long. It will not be dislodged so easily. Hence, while Tigard suggests that the advent of artificial agents will "provide new conceptions of moral agency,"[46] I am more disposed to think that artificial agents will have to conform to established conceptions of moral agency. One core tenet of those established conceptions is that only *real* moral agents can bear *real* moral responsibility. There is no telling how old this idea is. What is new is the need to state it explicitly.

The real/artificial distinction is real, not artificial. After all, one does not need to ascribe a "soul" or "mind" to gold in order to insist on its difference from fool's gold. The emphasis on blaming or praising only agents who are really responsible thus fits with ordinary practices. We may err in determining what exhibits real agency—wrongly crediting an albatross with bringing good sailing weather, say—but once we discover the error of our ways, we stop holding the target in question responsible. To make this concrete, those of us who do not think that God exists do not fear His wrath or worry about what course of action He allegedly prescribes. Where ontology goes, normativity normally follows.

We created robots, so we know that they are not in charge (or are at least reflective of our choices—including the choice to build them in the first place). We are certainly free to coin new concepts to fit our new creations, but whether such concepts will have any purchase on the human imagination is another matter. I therefore surmise that emerging dilemmas about how to treat robots will force the real/artificial distinction to assume its place alongside the other pillars identified by moral psychology, like care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation.[47] What remains to be seen is whether artificiality is one consideration to weigh among others, or whether its presence renders all other moral standards inapplicable. We do not, after all, truly mourn the death of people who appear in movies, because we know full well that, no matter how convincing the acting is, those characters are fake. We are, by contrast, profoundly moved by a picture of a dead Syrian boy face down on a beach,[48] because it documents something *real.*

My main claim, in sum, is that "artificial moral responsibility" betokens moral responsibility to the same degree that a "fake orgasm" betokens an orgasm. Importantly, this separation of the real and the artificial remains unaffected by the prospect of robots that mimic our appearance and behaviors in ways that are practically indistinguishable. So, if one wishes to circumvent the widely held view that being *really* responsible matters, one must work to carefully disassemble the (often implicit) metaphysical assumptions subtending that view. The roots of those assumptions go deep. They might be flawed. But, if that is the case, then the flaw(s) must actually be identified and replaced with something better.

I agree that it is "because we wish to foster healthy moral development *in humans* and maintain moral concern for *each other* that we see reasons to treat some AI systems as if they were moral subjects and perhaps as something like moral agents" (emphasis in original).[49] Medieval persecution of inanimate objects gives a vivid indication of our "psychological and social need for rituals of justice."[50] There are limits to this, though. I suppose that, instead of switching off dangerous robots, we could build maximum-security prisons where robotic guards enforce "as if" sentences on robotic inmates who display "as if" boredom and/or remorse. This, however, seems like an expensive price to pay for soothing our evolved propensity for agency-detection and face-recognition. It is in this sense that addressing the ontology of robot responsibility is a mandatory task. Until and unless this is done, the artificial conceptions of morality currently toyed with by some ethicists will feel, well, artificial.

Notes

1. Herrmann E, Call J, Hernández-Lloreda MV, Hare B, Tomasello M. Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science* 2007;**317**(5843):1360–6.
2. For more on morality and the propensity to project faces (called pareidolia), see Friesen BK. *Moral Systems and the Evolution of Human Rights*. Springer: Dordrecht; 2015:28.
3. Stowers K, Leyva K, Hancock GM, Hancock PA. Life or death by robot? *Ergonomics in Design* 2016;**24**(3):18.
4. Bringsjord S. *What Robots Can and Can't Be*. Dordrecht: Springer; 1992:3–4.
5. Dalton-Brown S. The ethics of medical AI and the physician-patient relationship. *Cambridge Quarterly of Healthcare Ethics* 2020;**29**(1):116.
6. See note 4.
7. Allen C, Wallach W. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press; 2009.
8. 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.
9. Dennett DC. *The Intentional Stance*. Cambridge: MIT Press; 1987.
10. An ambitious attempt to explain the intentional stance's predictive success can be found in Dennett DC. Real patterns. *The Journal of Philosophy* 1991;**88**(1):27–51.
11. Fodor JA. *A Theory of Content and Other Essays*. Cambridge: MIT Press; 1994:6.
12. Fake nurse in Quebec discovered and fired—after 20 years on the job. *Montreal Gazette* 2019 Jun 1.
13. See note 12. Interestingly, if we go back to his actual writings, we find that the founder of pragmatism Charles Sanders Peirce invited us to "[c]onsider what effects, which *might conceivably* have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object" (emphasis added). Peirce CS. *The Essential Peirce*, Bloomington: Indiana University Press; 1992:132. Highlighting that a difference makes no practical difference is thus persuasive only if it can be shown that the difference at hand cannot possibly make any practical difference in any conceivable future. Clearly, no robotic duplicate can meet this more demanding standard.
14. See note 8.
15. See note 8.
16. Asimov I. *The Complete Robot*. Garden City: Doubleday; 1982:3–6.
17. See note 8.
18. See note 8.
19. Champagne M. Disjunctivism and the ethics of disbelief. *Philosophical Papers* 2015;**44**(2):139–63.
20. Strawson PF. *Freedom and Resentment and Other Essays*. London: Routledge; 2008:23.
21. Himma K. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology* 2009;**11**(1):19–29.
22. See note 7.
23. See note 8.
24. Champagne M. *Consciousness and the Philosophy of Signs: How Peircean Semiotics Combines Phenomenal Qualia and Practical Effects*. Cham: Springer; 2018.
25. See note 8.
26. See note 8.
27. Berman PS. Rats, pigs, and statues on trial: The creation of cultural narratives in the prosecution of animals and inanimate objects. *New York University Law Review* 1994;**69**:288–326.
28. Sparrow R. Killer robots. *Journal of Applied Philosophy* 2007;**24**(1):10.
29. Champagne M, Tonkens R. Bridging the responsibility gap in automated warfare. *Philosophy and Technology* 2015;**28**(1):126.
30. See Champagne M. Axiomatizing umwelt normativity. *Sign Systems Studies* 2011;**39**(1):23.

31. See note 8.
32. The episode, titled The measure of a man, is the ninth episode of the series' second season and originally aired on February 13, 1989.
33. Hart HLA. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Oxford University Press; 2008:210–30.
34. Shoemaker D. Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics* 2011;**121**(3):602–32.
35. Ortenzi V, Controzzi M, Cini F, Leitner J, Bianchi M, Roa MA, Corke P. Robotic manipulation and the role of the task in the metric of success. *Nature Machine Intelligence* 2019;**1**(8):343.
36. Edmonds M, Gao F, Liu H, Xie X, Qi S, Rothrock B, Zhu Y, Wu YN, Lu H, Zhu SC. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics* 2019;**4**(37): eaay4663.
37. Brandom RB. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge: Harvard University Press; 1994:87.
38. Williams G. Responsibility as a virtue. *Ethical Theory and Moral Practice* 2008;**11**(4):460.
39. See note 36, Edmonds et al. 2019:9.
40. The shorter life spans of replicants, which is a much more important difference, would also have to match ours.
41. I am borrowing this metaphor from Jordan Peterson. See Champagne M. *Myth, Meaning, and Antifragile Individualism: On the Ideas of Jordan Peterson*. Exeter: Imprint Academic; 2020.
42. See note 8.
43. Bartels DM, Pizarro DA. The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition* 2011;**121**(1):154–61.
44. See note 8.
45. See note 8.
46. See note 8.
47. Graham J, Haidt J, Koleva S, Motyl M, Iyer R, Wojcik SP, Ditto PH. Moral foundations theory. *Advances in Experimental Social Psychology* 2013;**47**:55–130.
48. Henley J, Grant H, Elgot J, McVeigh K, O'Carroll L. Britons rally to help people fleeing war and terror in Middle East. *The Guardian*; 2015 Sep 3.
49. See note 8. For instance, it is easy to see how the Flesh Fairs from the 2001 Steven Spielberg movie A.I., where robots are destroyed "in a cross between a county fair, heavy metal concert, and WWF match or monster truck rally," could make one insensitive to violence toward humans. See Kreider T. A.I.: Artificial Intelligence. *Film Quarterly* 2002;**56**(2):36.
50. See note 27, Berman 1994:326.

Cite this article: Champagne, M. 2021. The Mandatory Ontology of Robot Responsibility. *Cambridge Quarterly of Healthcare Ethics* 30: 448–454, doi:10.1017/S0963180120000997