11. Master Copy	All track changes and comments.

Open Peer Commentary

Unpredictable robots elicit responsibility attributions

Matija Franklin[a], Edmond Awad[b], Hal Ashton[c], David Lagnado**[Q1]**[a]

[a] Experimental**[Q2]** Psychology Department, University College London, London, UK

[b] Economics Department, University of Exeter, Exeter, UK

[c] Computer Science Department, University College London, London, UK. matija.franklin@ucl.ac.uk; https://www.ucl.ac.uk/pals/research/experimental-psychology/person/matija-franklin/ e.awad@exeter.ac.uk; https://www.edmondawad.me ucabha5@ucl.ac.uk; https://algointent.com/ d.lagnado@ucl.ac.uk; https://www.ucl.ac.uk/pals/research/experimental-psychology/person/david-lagnado/

History : accepted : 2022-07-06

Copyright Line: Copyright © The Author(s), 2022. Published by Cambridge University Press

Abstract

Do people hold robots responsible for their actions? While Clark and Fischer present a useful framework for interpreting social robots, we argue that they fail to account for people's willingness to assign responsibility to robots in certain contexts, such as when a robot performs actions not predictable by its user or programmer.

Autonomous machines are increasingly used to perform tasks traditionally undertaken by humans. With little or no human insight, these machines make decisions that significantly impact people's lives. Clark and Fischer (2022) (C&F) argue that people conceive of social robots as depictions of social agents. They differentiate between the "base scene" – representing the physical materials the robot is made from, the "depictive scene" representing the robot's recognizable form along with an interpretive authority, and the "scene depicted" which either transports people into an imagined world inhabited by the robot or imports the robot's imagined character into the real world. We argue that this framework fails to account for people's willingness to assign responsibility to social robots (and AI more generally). Specifically, we argue that in a range of cases people assign some degree of responsibility to social robots, and do not shift all responsibility to the "authority" that uses the robot. These cases include robots that behave in novel ways not predictable by their users or programmers. We also argue that responsibility attribution is not a finite resource; thus users and robots can simultaneously be held responsible.

Recent work (Tobia, Nielsen, & Stremitzer, 2021) explores the question of who is held responsible for the actions of autonomous machines. Experimental evidence suggests that people are willing to attribute blame or praise to robots as agents in their own right (Ashton, Franklin, & Lagnado, 2022; Awad et al., 2020; Franklin, Awad, & Lagnado, 2021). As agents, autonomous machines are sometimes treated differently from humans. For example, people tend to hold humans accountable for their intentions while holding machines accountable for the outcomes of their actions (Hidalgo, Orghian, Canals, De Almeida, & Martin, 2021). Further, people ascribe more extreme intentions to humans while only ascribing narrow intentions to machines. This is a puzzle for the depiction framework because it shows that people are prepared to attribute responsibility to *depictions* of agents as well as to the depiction's authority.

C&F argue that attributing responsibility to a depiction's authority is intuitive for the ventriloquist's dummy or a limited social robot like Asimo. However, the examples they list in Table 2 concern those whose behavior is largely predictable, at least by the authority. Recent technological advances have produced social robots capable of generating original behavior not conceived even by their creators (Woodworth, Ferrari, Zosa, & Riek, 2018). Using machine learning methods, modern social robotics learn human preferences by observing human behavior in various contexts, developing adaptive robot behavior which is tailored to the user (Wilde, Kulić, & Smith, 2018). The mechanisms by which they reach their decisions are opaque, complex, and not directly encoded by the creator. We propose that such social robots are more likely to elicit

Perceived increases in machine autonomy come with increases in attributed responsibility toward those machines. First, higher machine autonomy is associated with intent inferences toward machines becoming more like humans (Banks, 2019). Thus research shows that when robots are described as autonomous, participants attribute responsibility to them nearly as much as they do to humans (Furlough, Stokes, & Gillan, 2021). Additionally, more autonomous technologies decrease the perceived amount of control that the authority has over them, which in turn decreases the credit the authority receives for positive outcomes (Jörling, Böhm, & Paluch, 2019). Similarly, drivers of manually controlled vehicles are deemed more responsible than the drivers of automated vehicles (McManus & Rutchick, 2019).

Furthermore, C&F's assertion that the creator of the depiction is responsible for the interpretation of their depictions relies on the fact that the depiction's behavior is predictable by the creator. The authors write: “We assume that Michelangelo was responsible not only for carving David, but for its interpretation as the biblical David” (target article, sect. X, para. X)**[Q3]**. But this argument fails for machines that behave unpredictably. When the painting “Edmond De Belamy,” generated by a deep learning algorithm, sold at an art auction for $432,500, many credited the machine (Christie's, 2018). This attribution to machine creativity goes beyond anecdotal evidence (Epstein, Levine, Rand, & Rahwan, 2020). Similarly, AlphaGo, in beating World Champion Go-player Lee Sedol, used novel strategies as adopted by human players (Chouard, 2016). Such novel moves prompted comments worldwide about machine creativity (McFarland, 2016), giving credit to AlphaGo rather than just DeepMind's team. While the DeepMind team intended AlphaGo to win the match, they did not envisage these novel moves.

Moreover, accounts of responsibility attribution should avoid committing the fixed-pie fallacy (Kaiserman, 2021) – the false assumption that there is a total amount of responsibility that can be allocated, or in other words, treating responsibility as a finite resource. The statement “*when Ben interacts with Asimo, he would assume that there are authorities responsible for what Asimo$_{char}$ actually does…*” (C&F~~target article~~, sect. 8.1~~X~~, para. 4~~X~~) hints at this error. People are willing to attribute responsibility to both autonomous machines and their users (e.g., a self-driving car and the driver; Awad et al., 2020).

There are also strong normative arguments that go against this fixed-pie fallacy. Some argue that neither the creators nor the operators of autonomous machines should bear sole responsibility (Sparrow, 2007). Others have drawn parallels between artificial intelligence and group agency – usually assigned to large corporations – as both are nonhuman goal-directed actors (List, 2021). Even in the case of recent fatal autonomous car crashes, attribution of legal responsibility to the car's manufacturer has not proved as straightforward as C&F's model would predict (De Jong, 2020).

C&F present an insightful framework to cover predictable and pre-programmed social robots. Here we have argued that more intelligent, autonomous, and thus, unpredictable social robots exist today. People are willing to attribute responsibility to such robots for their mistakes (Ashton et al., 2022; Awad et al., 2020; Franklin, Ashton, Awad, & Lagnado, 2022). Further, for more anthropomorphized social robots, research suggests that people are even willing to attribute experiential mental states (Fiala, Arico, & Nichols, 2014). The framework thus needs to be extended to handle the more intelligent robots currently being produced, and normative theories in philosophy and law suggesting that social robots may need to share social responsibility.

Financial support

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest

None.

References

Ashton, H., Franklin, M., & Lagnado, D. (2022). Testing a definition of intent for AI in a legal setting. Unpublished manuscript.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., … **Rahwan, I.** (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, **4**(2), 134–143.

Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior*, **90**, 363–371.

Chouard, T. (2016). The go files: AI computer wraps up 4–1 victory against human champion. *Nat*

Christie's (2018). Is artificial intelligence set to become art's next medium? [Blog post]. Retrieved from https://www.christies.com/features/a-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

Clark, H. H., & Fischer, K. (2022). Social robots as depictions of social agents. *Behavioral and Brain Sciences*, 1–33.**[Q5]**

De Jong, R. (2020). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. *Science and Engineering Ethics*, **26**(2), 727–735.

Epstein, Z., Levine, S., Rand, D. G., & Rahwan, I. (2020). Who gets credit for AI-generated art?. *iScience*, **23**(9), 101515.

Fiala, B., Arico, A., & Nichols, S. (2014). You, robot.**[Q6]**

Franklin, M., Ashton, H., Awad, E., & Lagnado, D. (2022). Causal Framework of Artificial Autonomous Agent Responsibility. In Proceedings of 5th AAAI/ACM Conference on AI, Ethics, and Society (AIES ’22) .**[Q7]**

Franklin, M., Awad, E., & Lagnado, D. (2021). Blaming automated vehicles in difficult situations. *iScience*, **24**(4), 102252.

Furlough, C., Stokes, T., & Gillan, D. J. (2021). Attributing blame to robots: I. The influence of robot autonomy. *Human Factors*, **63**(4), 592–602.

Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., & Martin, N. (2021). How humans judge machines. MIT Press.

Jörling, M., Böhm, R., & Paluch, S. (2019). Service robots: Drivers of perceived responsibility for service outcomes. *Journal of Service Research*, **22**(4), 404–420.

Kaiserman, A. (2021). Responsibility and the “pie fallacy”. *Philosophical Studies*, **178**(11), 3597–3616.

~~Lim, D. (2018). AI & IP: Innovation & creativity in an age of accelerated change. *Akron Law Review*, **52**, 813.~~**[Q8]**

List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology*, **34**(4), 1213–1242.

McFarland, M. (2016). What AlphaGo's sly move says about machine creativity. The Washington Post, retrieved from washingtonpost.com/news/innovations/wp/2016/03/15/what-alphagos-sly-move-says-about-machine-creativity/

McManus, R. M., & Rutchick, A. M. (2019). Autonomous vehicles and the attribution of moral responsibility. *Social Psychological and Personality Science*, **10**(3), 345–352.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, **24**(1), 62–77.

Tobia, K., Nielsen, A., & Stremitzer, A. (2021). When does physician use of AI increase liability?. *Journal of Nuclear Medicine*, **62**(1), 17–21.

Wilde, N., Kulić, D., & Smith, S. L. (2018). Learning User Preferences in Robot Motion Planning through Interaction. 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 619–626). IEEE.

Woodworth, B., Ferrari, F., Zosa, T. E., & Riek, L. D. (2018). Preference Learning in Assistive Robotics: Observational Repeated Inverse Reinforcement Learning. Machine Learning for Healthcare Conference (pp. 420–439). PMLR.