

Jakob Lucko
Dr. Sorgner
PH 323-1
April 24, 2024

AI's Character, Consciousness, and Culpability:
Can AI be Morally Responsible?

Introduction

Many dilemmas facing the AI debates surround the questions of personhood and consequently accountability. The idea that AI should not be held accountable for any of its actions arrives with the notion that AI is not a person; rather a machine *programed* to complete a task. However, if the deciding factor in accountability is agency and consciousness— like many philosophers bestow on humans— then I argue *if* AI can be categorized as a person based on the classical and contemporary ideas of consciousness and free will, then robots should be held morally accountable for their actions as causal agents.

This essay seeks to incorporate findings on agency, free will, and consciousness made by philosophers like Frankfurt, Strawson, Nagel, Dennett and others to buttress my argument that AI should be held morally accountable. I will first attempt to prove that AI could be considered a person, characterize AI's will and reasoning, then establish if AI has a conscious, and finally I will discuss the necessity to believe that AI has a conscious. By the end of the essay, I hope to prove moral accountability is necessary as AI continues to grow.

AI and Personhood

Using Harry Frankfurt as the framework for my contentions surrounding personhood and desires, I will first layout a vital preliminary notion about the term *persons*.¹ Although it is used as the singular version of people, humans use it because we think the qualities that define a person are unique to humans; even though being a person is not necessarily tied with our species, and other species could even consider themselves persons. But Frankfurt separates humans from animals not by personhood, but by how we make our choices. Animals have a desire and make a choice, while humans have second order desires which allows us to want to want things or not to want them while also thinking to want to think differently. Frankfurt defines the difference in simple terms, “that someone has a first-order desire when he wants to do or not to do such-and-such, and that he has a second-order desire when he wants to have or not to have a certain desire of the first order.”² In other words when an agent has first-order desires, they simply want a desire (*A* desires/wants *X*), but this does not indicate how much relative strength *A*’s desire to *X* and if this desire will play a role in what *A* will actually do.³ And the second-order desire— as he reconstructs— is the effective desire that moves (or would move) an agent all the way to action.⁴ Insofar that what drives the agent to act, despite any conflicting desire or intuition, is the agent’s will, then when applied to AI, we can successfully say that AI has a will.⁵ In other words, on the scale of agency and desire, according to Frankfurt, AI surpasses first-level desires.⁶ However, AI

¹ Harry Frankfurt, “Freedom of the Will and the Concept of a Person.” *The Journal of Philosophy* 68, no. 1. (1971), p. 7.

² Ibid., (1971).

³ Frankfurt, “Freedom of the Will,” (1971), p. 8. I will also include a concise summation of his point: “For even though someone may have a settled intention to do *X*, he may nonetheless do something else instead of doing *X* because, despite his intention, his desire to do *X* proves to be weaker or less effective than some conflicting desire.”

⁴ Ibid., (1971).

⁵ I assert that AI appears to *will* for their program to come true.

⁶ This is curious because AI seems to bypass first order desires and goes straight to second order desires, or the desires that cause us to act. However, the question is if AI knows that they are desiring this.

does not qualify in personhood quite yet because Frankfurt's addition of second order volitions— which is the ability to reflect on one's desires and make the desire to fulfill one's will a first-level desire— opens up the conversation to whether AI has the ability to desire its own will and for Frankfurt, the genuine threshold for what qualifies personhood.⁷

AI as a Wanton?

It would seem that Frankfurt tinkers on the premise that an agent needs a conscious in order to have second-order volitions and reflect on its will to make sure their desires align with their effective will and consequently to be considered a person.⁸ However, he hypothesizes a creature that has first-order desires and may or may not have second order desires but doesn't have second-order volitions. Calling it the wanton, Frankfurt characterizes this as something who does not care about his will, rather, "His desires move him to do certain things, without its being true of him either that he wants to be moved by those desires or that he prefers to be moved by other desires."⁹ But could AI be categorized as a wanton if what distinguishes wantons from rational humans is their inability to be concerned with the desirability of their desires? AI has a will and appears to desire its will completed, but the lack of consciousness to understand if it desires to know if its desires align with its genuine will, is the limit of AI's ability. AI's lack of a conscious would prove that AI cannot be held morally responsible for anything it does, not because it's full wanton, but because it does not have the nature to understand its desire and volitions.

⁷ Frankfurt, "Freedom of the Will," (1971), p. 10.

⁸ Ibid., (1971).

⁹ Frankfurt, "Freedom of the Will," (1971), p. 11.

The Need for Consciousness

Joshua Shepherd conducted an experiment and he found that people are far more likely judge an agent as free, blame or praiseworthy, and responsible if they agent is conscious rather if the agent is acting non-consciously.¹⁰ The study proves the widely accepted theory that consciousness is necessary for free will and to an extent moral responsibility. This is because if an agent has consciousness and understands what desires align with his effective will and still chose to act immorally, regardless of if they could have done otherwise, the agent is morally responsible.¹¹ This leaves AI utterly unrestricted from moral responsibility even though it has a will and technically desires to complete their will; yet the missing piece is their awareness that and their desire for their will controlling their desires. However, more nuanced ideas which critique Frankfurt offer some more room to still find moral accountability for AI.

Is AI's Reasoning Enough?

Daniel Dennett proposes the idea that if automated robots have the ability to rationally self-reflect, then, “They would be equipped to react appropriately when we represent reasons to them. Isn’t that what freedom hinges upon, whether or not it amounts to consciousness?”¹² And if the agent— in this case the automated robot— has freedom, we essentially skip over the criterium of consciousness in the equation of moral responsibility and rely on the free will

¹⁰ Eddy Nahmias, Corey Hill Allen, and Bradley Loveall, “When Do Robots Have Free Will?: Exploring the Relationships between (Attributions of) Consciousness and Free Will.” In *Free Will, Causality, and Neuroscience*, edited by Bernard Feltz, Marcus Missal, and Andrew Sims, (Brill, 2020), p. 60.

¹¹ Harry Frankfurt, “Alternate Possibilities and Moral Responsibility.” *The Journal of Philosophy* 66, no. 23 (1969).

¹² Daniel C. Dennett, *Elbow Room: Free Will Worth Wanting* (Oxford University Press, 1985), p. 43. Quoted by Nahmias, Allen and Loveall (2020), p. 63.

requirement.¹³ Many quality of will philosophers¹⁴ derive their theories from Peter Strawson who argues “That freedom and responsibility are grounded in our reactive attitudes, such as indignation, gratitude, and guilt, that we express in interpersonal relationships.”¹⁵ In other words if an agent can express or experience certain moral emotions, then we attach responsible agency only to those who we think are also feeling or expressing the relevant emotions, then we can characterize them as (im)moral.¹⁶

Does AI Care?

Given this approach on free will and moral accountability, the issue with AI is, if we think they can feel guilt, indignation, or gratitude and the closer AI gets to feeling (or perfectly faking) these emotions that give humans the ability to *feel* moral, the closer AI gets to being fully morally accountable for its actions. However, Frankfurt modifies his stance and claims that in order to have freedom an agent must *care* about what is important to them.¹⁷ Again he expresses the need for reflection over the agent’s desires, but this emphasis on the importance of caring about one’s desires seems to diminish Strawson’s case.¹⁸ But according to Chandra Sripada, caring is an intrinsic foundational motive for any desire yet is also a finely tuned emotional

¹³ I will note that in the field there is known gap between consciousness and free will that Gregg Caruso says must be filled if we want to evaluate free will. The Nahmias, Hill Allen and Loveall article says this may derive from the intuitive connection of the two. Nevertheless, I am not taking advantage of this gap, but elaborating on the quality of will theorist claiming that claim consciousness is not required for free will and responsible actions. G.D Caruso, Consciousness, Free will, and Moral Responsibility in *The Routledge Handbook of Consciousness*, (New York: Routledge, 2016). p. 75-78. From Nahmias, Allen and Loveall (2020), p. 59.

¹⁴ Nahmias, Allen and Loveall (2020), p. 64, cites Arpaly (2003); Smith (2005); Sher (2009); Buss (2012); and Sripada (2016).

¹⁵ Ibid., (2016).

¹⁶ If AI has the capacity to fake these emotions, I will go as far to say that humans are utterly tricked into thinking they are moral or immoral entities given that we bestow moral responsibility based on if we think the other agent is capable of feeling these Strawsonian emotions.

¹⁷ Ibid., (2016).

¹⁸ I assert this because Strawson although makes a strong claim—that we can rely on if agents and objects express or can express these emotions, they can be morally accountable—Frankfurt’s rebuttal that *to care* entails that we still use our second order volitions would deem my use of Strawson’s argument invalid.

response.¹⁹ However Sripada neglects to include— as the Nahmias, Hill Allen and Loveall article points out— in order to care, the agent must already have phenomenal consciousness.²⁰ Consciousness is still a requirement for free will and moral accountability.²¹ So, it would seem that in either scenario, AI could still not be held morally accountable given that consciousness is the deciding factor.

Consciousness As the Deciding Factor

The Nahmias, Allen and Loveall group performed two experiments testing the relationship among free will, conscious sensation, Strawsonian emotions, consciousness manipulation, and basic emotions and found conclusive results that support the arguments above regarding the differences between how a humanoid robot's conscious is perceived compared to a human's.²² “These results suggest...the ability to feel emotions and have things actually matter to the individual is important in Free Will attributions, yet the ability to have conscious sensations specifically plays no significant role.”²³ Even if our AI has the capacity to react when slapped in the face and feel pain, it would require a Strawsonian emotional capacity to reach the type of freedom that would deem it morally responsible given that once again, it would appear that most people and theorists believe that consciences is crucial for free will and that robots do not have either.²⁴ AI's non-conscious state is also backed up by the *MIT Research Review*, as Grace Huckins states “According to *IIT*, conventional computer systems, and thus current-day

¹⁹ Nahmias, Allen and Loveall (2020), p. 65-67.

²⁰ The Nahmias, Allen and Loveall group says this is most likely because Sripada cites Shoemaker many times and it would be understood that emotions require phenomenal consciousness. (2020), p. 65.

²¹ Ibid., (2020).

²² Nahmias, Allen and Loveall (2020), p. 68-76.

²³ Nahmias, Allen and Loveall (2020), p. 76.

²⁴ Nahmias, Allen and Loveall (2020), p. 77.

AI, can never be conscious —they don't have the right causal structure.”²⁵ However, if AI lacks the consciousness this means AI cannot feel anything, no matter how intelligent it becomes.²⁶ This is a curious matter because we treat AI like it has a personality, opinions and a distinct perspective; like it has a consciousness.²⁷ Robert Long shines a light on the simple program called ELIZA, built in the 1960s to simulate talk therapy which confused many people into thinking AI is capable of feeling, caring, and thought. But the idea is, as AI becomes more convincing with its perception of conciseness via realistic conversations or seemingly involuntary acts, humans will blur the lines of AI's intelligence and its illusionary consciousness.²⁸

Should Humans *Think* AI has Consciousness?

The problem now lies when humans think AI is conscious when they are not; we cannot bestow any moral associations onto AI because intrinsically, automated machines do not have free will and is not conscious.²⁹ If AI truly does not care about anything because they lack a conscious, yet humans are disillusioned into thinking they do, a point of conflict will arise when private companies like Google or Open AI furthers developing machines with consciousness because as Rufin VanRullen fears, these companies won't “welcome the ethical quandaries that a

²⁵ Grace Huckins, “Minds of Machines: The Great Ai Consciousness Conundrum,” (MIT Technology Review: October 16, 2023).

²⁶ Nahmias, Allen and Loveall 67.

²⁷ Grace Huckins, “Minds of Machines,” (2023). cites Robert Long, philosophy fellow, Center for AI Safety in San Francisco.

²⁸ Grace Huckins, “Minds of Machines,” (2023). cites Robert Long.

²⁹ Grace Huckins makes a curious reference to Star Trek and other Sci-fi movies about robots and claims, “no matter how intelligent or complex their behavior, they do not come across as free and autonomous until they seem to care about what happens to them (and perhaps others).” A fitting comment for how perceived emotions plays into having conscious. “Minds of Machines,” (2023).

conscious system would introduce.”³⁰ These ethical quandaries are presented above regarding the tie between consciousness, freedom, and moral responsibility set forth by Frankfurt, Sripada, Strawson, and others. AI could be acting immoral; and unknowingly, humans would still uphold non-morally responsible criteria to a now-conscious entity. However, if the fear of AI is if when it becomes conscious, there are three options: The first, as supported by Eric Schwitzgebel and Mara Garza, says that if it is unclear whether an AI has consciousness, it should not be built.³¹ The second, is to treat all AI like it is conscious because the time it turns conscious is unpredictable. And the third is to rethink what consciousness means in attempt to remove ambiguity from the gray zone.³²

Rethinking Consciousness

As stated previously, the concept of consciousness is widely debated, yet very little is known. However, in order to understand current AI and its next leap into consciousness, the concept should be explored more.³³ Thomas Nagel suggests that since conscious experience exists in other life forms, there is something it is to be like that organism.³⁴ Deeming this essence the subjective character of experience, Nagel affirms that even if we knew everything about an agent— in Nagel’s case, a bat— we would still lack insight into its unique quality of conscious

³⁰ Grace Huckins, “Minds of Machines,” (2023). Also notes that VanRullen relies on publicly funded visible projects like his own to make sure the grey zone of AI’s consciousness is publicly shared and known unlike if the private sector establishes a conscious AI, the public would most likely not know, which he finds worrisome.

³¹ Ibid., (2023). Also note that both theorists claim this is unlikely especially if the technology is profitable.

³² The grey zone refers to the debated topic of if AI can truly exhibit consciousness or conscious-like qualities.

³³ I find the fields’ knowledge on consciousness very underdeveloped, especially for a concept that is so extensively referenced and explored. As far as my research goes, it appears that philosophers have widely accepted that the conscious is an unknown, intrinsic, human-owned facet of the mind.

³⁴ Thomas Nagel, “What Is It Like to Be a Bat?” *The Philosophical Review* 83, no. 4, (Duke University Press: 1974), p. 436. Nagel says in regard to robots in footnote 2. “Perhaps anything complex enough to behave like a person would have experiences. But that, if true, is a fact which cannot be discovered merely by analyzing the concept of experience.”

experience given that an outside bystander cannot fully know another's subjective private experiences or facets of the mind.³⁵ Challenging of the conventional idea of consciousness, Nagel offers a new light on AI's consciousness. If we follow the Nagelian route, AI could have its own subjective character that we are unaware of like the bat's. When investigating AI, we are merely inspecting its programs and superficial mechanical engineering, but I suggest if humans do not know what it *is* to be an AI, then I argue we should be treating AI as conscious beings already. Yes, humans may have the upper hand in that we have a thorough understanding of AI than AI itself, but humans have no idea what it is like to be an AI.

Conclusion

The arguments against AI having a conscious seem to outweigh the contrary, however, bestowing consciousness on AI before it becomes conscious would prepare humans for how to assign moral responsibility and draw a hard line in the sand to rid the gray zone of any ambiguity that theorist fear. Although it is accepted that no current consciences in AI, Strawson, Nagel, and Frankfurt prove that autonomy (a goal for many AI developers), consciousness (AI's ability to appear that they care), and second-order volition (AI's ability to self-recognize its will) are all prerequisites for moral responsibility. And all not so far off in the future. This is why the sooner we acknowledge AI having a conscious with agency and moral responsibility, the better prepared we will be when they finally do.

Word Count: 2119

³⁵ Nagel, "What Is It Like to Be a Bat?" (1974), p. 437- 443.

Bibliography

- Bennett, Jonathan, 2008, “Accountability (II)”, in *Free Will and Reactive Attitudes: Perspectives on P.F. Strawson’s “Freedom and Resentment”*, Michael McKenna and Paul Russell (eds.), Burlington, VT: Ashgate Publishing Company, pp. 47–68; reprinted London: Routledge, 2016.
- Caruso, G.D. Consciousness, Free will, and Moral Responsibility. In R.J. Gennaro (ed.), *The Routledge Handbook of Consciousness*, New York: Routledge, 2016. p. 78–90
- Dennett, Daniel C. *Elbow room: Free will worth wanting*. Oxford University Press, 1985.
- Frankfurt, Harry G. “Freedom of the Will and the Concept of a Person.” *The Journal of Philosophy* 68, no. 1 (1971): 5–20. <https://doi.org/10.2307/2024717>.
- Frankfurt, Harry G. “Alternate Possibilities and Moral Responsibility.” *The Journal of Philosophy* 66, no. 23 (1969): 829–39. <https://doi.org/10.2307/2023833>.
- Huckins, Grace. “Minds of Machines: The Great Ai Consciousness Conundrum.” *MIT Technology Review*, October 16, 2023. <https://www.technologyreview.com/2023/10/16/1081149/ai-consciousness-conundrum/#:~:text=According%20to%20IIT%2C%20conventional%20computer,have%20the%20right%20causal%20structure>.
- Nagel, Thomas. “What Is It Like to Be a Bat?” *The Philosophical Review* 83, no. 4 (1974): 435–50. <https://doi.org/10.2307/2183914>.
- Nahmias, Eddy, Corey Hill Allen, and Bradley Loveall. “When Do Robots Have Free Will?: Exploring the Relationships between (Attributions of) Consciousness and Free Will.” In *Free Will, Causality, and Neuroscience*, edited by Bernard Feltz, Marcus Missal, and Andrew Sims, 338:57–80. Brill, 2020. <http://www.jstor.org/stable/10.1163/j.ctvrk31x.8>.
- Van Inwagen, Peter. “Moral Responsibility, Determinism, and the Ability to Do Otherwise.” *The Journal of Ethics* 3, no. 4 (1999): 341–50. <http://www.jstor.org/stable/25115624>.
- Van Inwagen, Peter van. *Notes* 22, no. 4 (1988): 609–18. <https://doi.org/10.2307/2215460>.