Grahm Tuohy-Gaydos
Department of Philosophy
Ethics of Artificial Intelligence and Digital Technology
Lewis Williams
February 28th, 2024

How Ought the Moral and/or Legal Responsibility for Actions Performed by AI/Robots be Assigned?

To ease traffic, a city introduces AI-powered, reinforcement-trained stoplights capable of adapting to congestion patterns across the entire urban area. At one stop, however, every light turns green, and the ensuing pileup causes dozens of deaths. As the community reels, the question becomes singular: who is to blame for the tragedy? Was it the drivers' fault, the city council's, or the programmers behind the stoplights?

As the above scenario implies, the question of responsibility for AI decision-making is a challenging one, especially as these technologies have begun to integrate with legacy systems. As such, we stand at a critical crossroads: how can we assign legal and moral responsibility for the actions of artificial intelligence and automated systems? This essay will demonstrate how these systems, more than just producing a 'responsibility gap', create a 'vacuum of moral responsibility' that will obfuscate any attempt to assign blame or determine the central moral agent within any incident. As such, responsibility will likely be allocated through legal means as a substitute for deriving true moral accountability.

In pursuit of this finding, this essay will begin with a consideration of the larger issue of responsibility in AI systems. The 2nd section of this analysis will develop the concept of a 'moral vacuum' as an extension of the existing concept of the responsibility gap. Finally, the third section will consider the vacuum's effects on the determination and assignment of obligation and accountability and the manner by which it magnifies the importance of legal responsibility. This will in turn provide evidence for the elevated importance of legal determination as a substitute for true moral accountability.

The Issue of Responsibility in AI Systems

When trying to determine the culpability of AI in any instance, it is important to view accountability as an act of responsibility, which is by definition hard to define. It can refer to blame, liability, or obligation and as a result, I have chosen to use the synonyms of responsibility interchangeably rather than attempt to differentiate between each. More substantively, I have chosen to broadly consider responsibility as made up of two separate approaches: the legal and the moral.

Legal Responsibility. Mingle and Reagan note that, "legal responsibility is virtually synonymous with liability; The two are often used interchangeably," (Mingle and Reagan, 1983, p. p. 115). Legal responsibility is concerned with the question of blame and punishment. More broadly, it can be understood as one centred around the capability of enforcement through a system of broader obligations. As Johnson defines it, "[It is] an obligation that outside parties are prepared to enforce in a regular way, using publicly available procedures to determine the fact of violation and the way violations will be handled," (Johnson, 1975, p. 332).

Moral Responsibility. Comparatively, moral responsibility is far harder to define. Given its connection to larger precepts which may or may not be reflected in civil code, establishing a clear basis for moral responsibility can pose a challenge. In the context of this paper, it can be understood as an ethical imperative, expectation, or requirement which directly emerges from one's status as a moral agent. For example, it may be one's moral responsibility to stop for a broken-down car on an empty road, but not necessarily a requirement; it is rooted in a deeper basis of empathy and connection which define moral agency.

Differentiating between legal and moral responsibility is in many cases, a challenge—where is the line between the two? Does legal responsibility imply a moral responsibility? I have chosen to avoid these larger conversations given the limitations of this forum. That said, there is use in considering the relationship between the two further, especially in light of the role each will play in assigning responsibility for AI use.

With the legal and moral basis of the argument now established, we can return to the case study in the introduction to assign moral responsibility while setting aside legal responsibility for the time being.

The Programmer. If the system is autonomous and operating outside the programmers' intent, then they cannot be held to account from the moral perspective given that they could not be expected to have prevented or otherwise stopped such an incident.

The Buyer/User. Grouping the city council and the drivers together, neither group could be held responsible unless either were fully aware of the capabilities of the system or in some way abused it beyond its capacities. In the absence of any act which superseded the AI or an awareness of its ability to commit the wrongful act in question, neither group can arguably be blamed.

The AI. While technically the propagator of the inciting incident, the artificial intelligence cannot be fully blamed. As Véliz notes, "When moral agents hurt others, we can blame them for their bad intentions or their neglect. In contrast, we do not feel moral outrage against algorithms because they could not have acted otherwise, given their design and input, and they do not have intentions—they do not feel ill will or contempt," (Véliz, 2021, p. 7). As has been convincingly demonstrated in the literature (see Véliz 2021) and among the larger scientific community, AI's lack of moral agency necessarily excuses it

> from reapproach; without sentience, it cannot exhibit moral reasoning, and without the necessary empathy and understanding, it cannot act as a true agent.

As such, we are left at an impasse. Each actor, barring extenuating circumstances, cannot be held responsible, even though the inciting event (the horrific accident) necessitates a response. By attempting to ascertain *moral* accountability, we have frustratingly found there to be none.

Responsibility Gap or Responsibility Vacuum?

The type of situation described above has often been termed as a moral responsibility gap which emerges from the autonomous nature of the system. Matthias, defining the term in 2004, describes it as,

> An increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume the responsibility for them (Matthias, 2004, p. 177).

Santoni De Sio and Mecacci, expanding on this analysis, identify four types of responsibility gaps: culpability, active responsibility, and moral and political accountability. The culpability gap results from an expanded set of legitimate excuses for wrongdoing due to AI; the active responsibility gap results from a lack of "sufficient awareness" for personal responsibilities surrounding AI safety; and moral and public accountability gaps result from the black box effects of these systems which make outcomes challenging to explain and the work of developers harder to scrutinise (Santoni De Sio and Mecacci, 2021, pp. 1059-1061). These four terms are useful indicators of the challenges produced by these systems and the moral gap that naturally emerges between the programmer/company and the AI itself.

The problem is that while this model is a useful descriptor of a critical problem in AI, it terminologically implies that moral responsibility continues to exist, but is simply lost 'in-between'. Expanding on Santoni De Sio and Mecacci's analysis, the nature of these four responsibility gaps extends the concept beyond simply the manufacturer-product dynamic into a vast network of intersecting obligations which is characterised by an *absence* of accountability and answerability. As such, AI more accurately produces a *vacuum of moral responsibility*. It is not simply a challenge of determining culpability, but one of an inability to even appraise responsibility in the first place; justice cannot be served, and no agent can be held to account.

Assigning Blame in a Moral Vacuum

Part of the challenge presented by this vacuum is that there are few equivalent examples from which to draw a solution for assigning moral blame. The one potentially insightful case study may be that of the insanity plea (NGRI). While legally and morally sound, NGRI produces a similar challenge—a crime has still been committed, but its circumstances mean that responsibility dissipates into a vacuum. Importantly, the insanity defence is one of guilt—the party admits to the given crime—but is absolved of responsibility in the process *from a legal perspective*. Since "there is no just punishment without desert and no desert without responsibility." (Morse, p. 783), the lack of moral agency and self-autonomy necessarily means that punishment cannot be deliberated. Critically, this produces a situation in which legal responsibility is assigned—the defendant is guilty—but moral accountability is not provided. From a legal perspective, the offending party has been effectively held to account, but by virtue of the plea itself, the criminal cannot be assigned blame for their otherwise immoral act. This is in large part the pretext for much of the consternation surrounding NGRI, even though it has existed in the law for thousands of years—we remain frustrated that the party is receiving leniency in the face of overwhelming evidence. Legal liability, a form of accountability in and of itself, can often feel hollow in the vacuum of moral accountability.

This is useful to consider in the context of AI. While not a one-to-one example, the insanity defence is indicative of the distance between moral and legal responsibility and how the two can differ. While we often understand that all that is morally required is not necessarily legally required, we fail to recognise that the inverse is true as well. In the moral vacuum produced by artificial intelligence, legal accountability will not similarly cease to exist—it will become *more* important. Programmers, fearing blame and consternation for technologies they are incapable of controlling, will become more likely to rely on documentation and terms of use to transfer blame onto the contractor or user. Legal definitions will place impositions on all users that otherwise may not exist in the moral dimension, creating a legally structured framework for ascertaining blame in cases in which ethics may otherwise fail. The moral element of these discussions will not cease to exist, but instead become secondary given the challenges posed by the vacuum of moral responsibility that AI produces. This process may make obligations and accountability feel hollow or overly simplistic, but in the vacuum, it represents the only potential outlet by which to consider a central concern in the age of artificial intelligence.

Conclusion

This analysis has served as a brief demonstration of how AI, more than just producing a 'responsibility gap', creates a vacuum of moral responsibility with cascading effects. Importantly, this vacuum will likely elevate the importance of legal responsibility, even if as the consideration of the insanity defence showed, it can often result in frustration. This essay has obviously taken a broader perspective, and further consideration of the differences between moral and legal accountability is of paramount necessity and may challenge elements of this argument. Similarly, further advancements in AI may influence these discussions. Should AI assume even limited moral agency, the situation shifts. Lastly, the actions of developers and the larger public will shape these conversations as well; Moral precepts are not simply integral elements of social life, but political constructs, and ethical guidelines or public outrage will continue to shape our perception of blame. Regardless, the issue of moral responsibility remains an important one, even if we currently sit at an impasse.

Bibliography

Ajoku, Chioma. n.d. "The Insanity Defense, Public Anger, and the Potential Impact on Attributions of Responsibility and Punishment."

Bonnie, Richard J. 1983. "The Moral Basis of the Insanity Defense." *American Bar Association Journal* 69 (2): 194–97.

Johnson, Conrad D. 1975. "Moral and Legal Obligation." *The Journal of Philosophy* 72 (12): 315–33. https://doi.org/10.2307/2025132.

Liao, S. Matthew. 2020. "The Moral Status and Rights of Artificial Intelligence." In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao, 0. Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0018.

Litwack, Thomas R. 1984. "The Moral Foundations of the Insanity Defense." *Criminal Justice Ethics* 3 (1): 12–19. https://doi.org/10.1080/0731129X.1984.9991739.

Matthias, Andreas. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

Mingle, John O., and Charles E. Reagan. 1983. "Legal Responsibility Versus Moral Responsibility: The Engineer's Dilemma." *Jurimetrics* 23 (2): 113–55.

Morse, Stephen J. n.d. "Excusing the Crazy: The Insanity Defense Reconsidered."

Mortazavi Azad, Seyedeh M., and A. Ramazani. 2023. "Smart Control of Traffic Lights Based on Traffic Density in the Multi-Intersection Network by Using Q Learning." *Discover Artificial Intelligence* 3 (1): 39. https://doi.org/10.1007/s44163-023-00087-z.

“Neuroscientist Who Studies How the Brain Learns Information Explains Why A.I. Would Be the ‘perfect Psychopath’ in an Executive Role.” n.d. Fortune. Accessed February 27, 2024. https://fortune.com/2023/07/31/why-ai-artificial-intelligence-perfect-psychopath-neuroscientist/.

Santoni De Sio, Filippo, and Giulio Mecacci. 2021. “Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them.” *Philosophy & Technology* 34 (4): 1057–84. https://doi.org/10.1007/s13347-021-00450-x.

Towers-Clark, Charles. n.d. “Without Human Values, What Stops AI From Acting As A Sociopath?” Forbes. Accessed February 27, 2024. https://www.forbes.com/sites/charlestowersclark/2023/11/21/without-human-values-what-stops-ai-from-acting-as-a-sociopath/.

Véliz, C. 2021. “Moral Zombies: Why Algorithms Are Not Moral Agents.” *AI and Society* 36. https://ora.ox.ac.uk/objects/uuid:d93dd717-b004-4280-861a-8f93a82eceb6.

“Will Artificial Intelligence Be Sociopathic? | Psychology Today United Kingdom.” n.d. Accessed February 27, 2024. https://www.psychologytoday.com/gb/blog/rich-encounters/202401/will-artificial-intelligence-be-sociopathic.