Science Fiction as a Guide for AI, Personhood, and Moral Consideration

Abigail Iris Backman-Daniels

Abstract

Science fiction has long been a source of provocative speculation that has influenced our conceptions of both the present and future. It can thus be argued that both science fiction and philosophy are united in a search for understanding, even though they may go about this search quite differently. This article explores some possible contributions of science fiction to moral philosophy, specifically regarding the question of moral consideration. Particular focus is given to the issue of Artificial Intelligence and personhood, and a number of case studies are used in this investigation. Isaac Asimov's Laws of Robotics and the short story *Cal* (1995), as well as the *Black Mirror* episodes "White Christmas" (2014) and "Be Right Back" (2013) are used to explore some science fiction narratives relevant to moral philosophy. In this exploration, the importance and relevance of science fiction to society, not only as a source of entertainment but also as having philosophical relevance, is highlighted. This article concludes that science fiction ought to be taken seriously and consulted as a guide for navigating AI, personhood, and moral consideration in the near future, given its unique capacity to explore such issues.

About the author

Abigail Iris Backman-Daniels (she/her) is currently a Master's student in the Philosophy Department. Her research interests include Artificial Intelligence, AI Ethics, AI-Human Relations, Future Studies and Future Societies, Fiction and Science Fiction, Science and the Technological Singularity. Her current thesis explores what it is like to engage with AI, and she sets out to pursue a PhD to explore the ethics surrounding AI-human relations with the intentions of establishing an ethical framework as well as practical guidelines. She completed an Honours (*cum laude*) in Philosophy with a focus on the interplay between science fiction and science fact as well as science fiction as a guide to possible futures, AI, and the technological singularity. She also completed an Honours (*cum laude*) in Ancient Cultures with a focus on Magic in Ancient Egypt, and has interests in Mythology, Middle Egyptian Hieroglyphs, and Linguistics. Outside academia, she is a Taekwon-Do black belt with provincial colours in three provinces, as well as national colours and three world champion titles.

1. Introduction

This article posits that the relationship between science fiction and philosophy becomes evident once one delves into the genre. It is for this very reason that I believe there should be continued philosophical research into and discussion of science fiction, given we are only as of recent decades unveiling the innate philosophical value which science fiction holds. To further this point, I would like to quote Manola Antonioli, who states:

> philosophy is also close to science-fiction in that one can write only about that which one knows badly, "at the edge of his knowledge" [*à la pointe de son savoir*], just as the science fiction writer always writes from the scientific knowledge of the present in the direction of a knowledge that we do not yet possess, or from this world in the direction of worlds that are possible but as yet unknown. (Antonioli, 1999, in Burton, 2015:12)

Despite science fiction regularly drawing on philosophy, and thought experiments in particular, it is not as common for philosophy to draw on science fiction. One possible reason for this could be the fear of having one's work discredited by other philosophers and academics (Tucker, 1996:534). There has, however, been considerable change in recent years with more philosophy discussing fiction as a whole. Despite this move, science fiction is still not regarded as philosophically serious work (Tucker, 1996:535–536). Ultimately, this ignores the philosophical importance of not only science fiction, but non-academic writing as a whole, and does an incredible disservice to the furthering of the knowledge basis. Even though there are science fiction works that take a philosophical perspective, there is not much philosophy discussed from a science fiction perspective. It is this which I hope to not only bring to attention but begin to bridge.

In this article, I draw heavily on both science fiction and moral philosophy in order to explore Artificial Intelligence (AI), personhood, and moral consideration. Science fiction — which the author and scientist Isaac Asimov described as "that branch of literature which deals with the reaction of human beings to changes in science and technology" (Blackford, 2017:8) — first arose in the 1600s, although only recognised as a distinct genre in the 1930s. As observed by Asimov in his definition, the genre formed as a literary response to the rapid industrialisation, scientific change, and technological innovations taking place (Blackford, 2017:5, 26). Moral philosophy, or ethics, refers to the branch of philosophy which explores the nature of ethics and morality; what is right and wrong, good and bad; one's moral intuitions; as well as ponders the question of how one ought to live and conduct oneself (Wolff, 2018:2, 4). The writing of this article was undertaken to ponder the question of AI's personhood and moral consideration as seen through the lens of science fiction. In this endeavour, the case studies of Isaac Asimov's Laws of Robotics and the short story *Cal* (1995), as well as the *Black Mirror* episodes "White Christmas" (Brooker, 2014) and "Be Right Back" (Brooker, 2013) are used because of their exploration of AI, personhood, and moral consideration. The chosen science fiction examples focus only on sentient and arguably conscious AI and on the negative impacts of denying AI personhood and moral consideration. Despite this limited scope of AI within the respective science fiction narratives, they are still of importance given that they begin the discussion surrounding AI, personhood, and moral consideration, as well as offer examples of how science fiction explores such moral questions. From this, the argument is made that science fiction has benefits to moral philosophy particularly as it pertains to the moral questions and dilemmas of AI. The case studies are used to illustrate how interactions and relationships with AI reflect how AI is conceived of and classified, and thus treated. This is undertaken to explore the consequences such classification and treatment of AI may have ethically, not only for AI, but also for humans as our treatment of other entities reflects our own moral standing and values. Through this, I build up the argument that science fiction can act as a guide to addressing such philosophical questions concerning the personhood and moral consideration of AI.

2. Science Fiction and Moral Philosophy

Science fiction is able to explore problems prominent in moral philosophy and explore possible solutions to said problems, as well as delve into possible issues which could arise with the introduction of new technological innovations (Mukerji, 2014:79). Fiction, particularly science fiction, enhances our moral understanding and empathy as it allows us to be confronted with the philosophical moral matter in a more personal manner such as witnessing other's perspectives from their own point of views and seeing how issues affect different individuals (Mukerji, 2014:79–80). Of specific interest is science fiction's introduction and exploration of philosophical issues pertaining to moral philosophy and the ethics of technological innovation (Mukerji, 2014:80). Yet, academic interest and engagement with the genre is quite novel (*ibid.*). In the words of Mukerji, "moral philosophers should watch sci-fi movies" (2014:81). This is because science fiction aids in complementing moral-philosophical research as it investigates existing and new issues as well as explores them and provides possible solutions. An example of this would be the familiar, age-old issues of agency, personhood, or consciousness, and the more contemporary issues of how AI challenges or forces the reshaping of our conceptions of these. Further, science fiction contributes to the field of moral philosophy by introducing philosophical issues to a wider audience who may not necessarily encounter these issues through the more traditional, academic channels. Science fiction also has the potential to foresee moral-philosophical issues not yet prominently raised by academics, namely those concerning robots and AI, before they become more common place. What makes science fiction particularly well suited for these endeavours is that it allows for the exploration of such ideas, principles, and concepts without the regulations of the natural world or limitations of current technological innovation.

Additionally, science fiction illustrates ethical systems in fully fledged out societies or worlds which gives deeper insight into what the nuanced "lived experience" within such ethical systems could look like and the consequences thereof on a social, political, and ethical level (*ibid.*; Blackford, 2017:75). Given that science fiction follows the narrative arc of a character and involves intense world-building, it allows readers/watchers to experience societal structuring, social classifications, ethical and legal systems, and technological advancements which one would not otherwise be privy to. In this sense, it has the benefits of thought experiments whereby different scenarios can be played out to see how principles or ideas can be implemented and what results they would yield. Unlike thought experiments, science fiction uniquely allows for more elaboration and nuance in this exploration which provides a fuller account. This is the case particularly for the exploration of principles and scenarios concerning AI, personhood and moral consideration as it is only in the nuanced, "lived experience" offered by science fiction that we begin to see the full extent of the results.

3. Case Studies

In this section, I outline the specific science fiction case studies and discuss the details of these to illustrate how the discussed issues can manifest in future societies. I discuss why these case studies are of moral philosophical interest and importance concerning the topic of AI and personhood. This will concern the treatment and classification of AI using select science fiction examples to demonstrate the possible outcomes and consequences of decisions regarding AI. Namely, I will look at the works of Isaac Asimov and the *Black Mirror* episodes "White Christmas" and "Be Right Back". I look at the programming and societal treatment of AI in the narratives as well as what this illustrates about AI, personhood, and the resulting moral considerations.

3.1. Asimov and the Three Laws of Robotics

The area of interest here is with Asimov's classification and treatment of robots[1], which is of interest given the current state of AI development. While currently existing AI is nowhere near as advanced as its science fiction counterparts, there are nevertheless ethical concerns explored in science fiction which we ought to have considered if such levels of technological advancement are ever reached. There are already issues seen in society resulting from a lack of boundaries drawn regarding AI in terms of how AI is classified, thought of, and treated. Asimov classifies and treats the robots as "lesser than" the humans, but this is questioned and problematised throughout different narrative arcs in his works. In Asimov's works, the robots generally serve a purely functional role, and their primary function is to bring

[1] Concerning the discussion of Asimov, "AI" and "robot" will be used interchangeably. In this context, the two are interchangeable. Asimov refers to them as "robots" but in our current understanding, he was in fact writing about embodied AI. This is clear given the sapience, sentience, and even consciousness displayed by his robots.

[2] The robots cannot break the laws in theory, given that it is part of their programming, but some of Asimov's works show the flaws in this as such laws are not definitive and can be broken or overridden in some instances.

about human happiness and comfort/ease of life, and secure their safety. In order to do this, the robots have the laws programmed into them which they cannot break[2]. These Laws of Robotics are as follows:

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. (Asimov, 2004:484–485).

What is of particular interest here is the prioritisation order of second and third law, which is constructed specifically so that the robot will prioritise human life above its own. Further, the robot's own protection or survival is secondary to that of a human's protection or survival. Asimov himself problematises this classification of AI as lesser than humans in certain short stories or parts of narratives whereby the robot's own wills and desires are prioritised over their human counterpart's. An example of this is the short story *Cal* (Asimov, 1995). This story follows Cal, a robot belonging to an author, who eventually develops an interest in writing and persists with trying to learn on his own. When his attempts are unsuccessful, and when he presents his nonsensical writing to his human master, his master calls in a programmer to install a dictionary in Cal's mind. Cal then writes words, but they are nonsensical sentences, so the programmer is called again to install grammar and so forth until Cal is capable of writing coherent sentences. After these new updates, Cal sets out writing his own stories but when his master reads them, he is worried about Cal's writing being better than his own, so he calls the programmer to undo the installations. Overhearing this conversation, Cal violates the first law by murdering his master in seeking out his own will and desire to be a writer, which he now values more than the law preventing harm to humans.

In this instance we see that if we essentially enslave an AI and treat them as lesser than us despite their sapience, sentience, or even consciousness — perceived or actual — there could be severe negative consequences. Such negative consequences involve individual resistance as seen with Cal or even in more widespread revolts where society could shut down or AI could "take over" and enslave us as is often feared with the technological singularity. Even if there are no such negative consequences which come to light, the classification and treatment of AI as lesser than holds as unethical. It can be seen in *Cal* that the denial of personhood and moral consideration of AI, when elements of personhood are displayed, not only leads to negative consequences for humanity but is unethical. When encountered with such narratives, we generally find the "subhuman" treatment of such entities as morally questionable at best and morally impermissible at worst. Here, the science fiction narrative allows us to not only encounter such moral dilemmas but also to grapple with how certain principles can play out if implemented as well as the short-term and long-term consequences of such. Asimov outlines a possible example of our future, one where AI is inferior and restricted by programming to prioritise humanity above all else. Through the case study of *Cal*, it is clear why this is not a desirable approach, given not only that such restrictive programming of AI's behaviour can be overridden, leading to negative consequences, but also because such an approach is unethical, given the perceived personhood of AI. This science fiction narrative proves to be philosophically interesting as it explores the consequences of essentially enslaving, restricting and controlling another sentient being. The story of *Cal* is philosophically important as its exploration of AI, personhood, and moral consideration can be used as a guide for navigating the future of AI as we ourselves grapple with the technological advancements of AI and the moral dilemmas it brings with it.

3.2. *Black Mirror*

The *Black Mirror* episode "White Christmas" (Brooker, 2014) explores complexities of personhood concerning the treatment of AI and humans. In this story, AI replicas of people are placed in egg-shaped objects called "cookies" to act as personal assistants for their human counterparts (Brooker, 2014). The cookies are digital replicas or simulations of a person's consciousness, accurate enough that the AI replicas believe themselves to be the original human. The cookies' purpose is to serve as a home automation device which will tailor everything in the house to the human's preferences.

Given that the cookie is a replica of the human, it will know these preferences exactly, from the time the blinds should open in the morning to wake the human, to when the coffee should be ready. At first, the cookie is reluctant to "cooperate" and serve its purpose, given that it believes itself to be the human. A representative from the company is present at the time of installation in order to "break in" the cookie. The cookies experience time differently as they can be programmed to experience extended periods of time while mere seconds elapse in the real world. This allows the representative from the company to simulate hours, days, weeks, and years in isolation for the cookie until they "break" and agree to cooperate by serving as home automated systems. It is shown that the cookies are capable of displaying pleasure and pain cues, as well as more complex emotional cues of an identity crisis when they are told they are in fact an AI replica. They experience distress when they are told they were created to serve as a home automation device, feel lonely/isolated in the simulated time which elapses, and are despondent at their fate.

The training and use of these cookies is rather ethically questionable, leaving us with a moral conundrum. Specifically, of concern is the ethical considerations of the treatment of AI in light of their own beliefs of their existence, insofar as they are capable of being conscious of their existence and believe themselves to be alive and human. Additionally, it is also interesting to look at how this differs from the ethical considerations of the treatment of humans in the same context. The AIs display cognitive and emotional intelligence; they display and arguably experience emotions; they have to be "broken in" to cooperate through the use of isolation torture, consisting of simulated years alone, or simulations to trick people into confessions. Here, the issue which is of interest is whether the cookie is actually conscious. It appears to be conscious, believes itself to be the person of which it is a replica, and it responds accordingly. It displays both pleasure and pain, happiness and suffering, as well as wills and desires. Such behaviours generally make one inclined to grant the being moral considerations given these displays of personhood and consciousness. The real-world example of this would be the general treatment of insects versus mammals[3]. It is generally morally permissible to kill insects, and we generally do so without any second thought. We generally consider insects to be of low to no consciousness as they do not display the familiar signs of pain and pleasure which we recognise. On the other hand, it is generally frowned upon to treat mammals such as dogs or elephants in the same way as they are considered to be of higher consciousness and display those signs of pain and pleasure which we recognise.

The area of interest in "White Christmas" (Brooker, 2014) is that the cookie is treated like a tool, albeit a sophisticated one, which is there to satisfy the human's wills and desires, not that of its own, despite displaying similar signs of pain and pleasure which we recognise. If we view and treat the AI as a mere copy despite its behaviours and responses equalling that of humans, we run the risk of unethical behaviour. If the AI is capable of displaying signs of suffering, what is to distinguish its suffering from a human being in the same situation displaying the same signs of suffering? According to Gomel, there is no ethical distinction (2011:349). Even if we cannot be certain whether the AI possesses consciousness, if the signs of suffering are sufficient, there cannot practically be any difference in moral consideration given to the suffering. When the AI is a close enough copy, the ethical boundary between AI and human becomes meaningless. Hence, there should be no ethical boundary which results in differing treatment of AI and humans. If we treat the AI as lesser than, as seen in the episode, what is to stop humans from being treated similarly under the justification of the abuser believing the human was AI or "lesser than"? Towards the end of the episode, this concern is directly explored when the same technology is used to elicit a confession from a prisoner. Concerning the ethical standpoint, if we are to classify AI as "lesser than" and treat them accordingly, we run the risk of immoral behaviour. Ultimately, the case study of "White Christmas" highlights the potential negative consequences of treating AI as "lesser than" in terms of how it is morally impermissible to deny the personhood and moral considerations of the AI when it displays signs of personhood in the form of suffering and happiness. This science fiction example is philosophically interesting as it explores the consequences of outright enslaving and abusing another sentient and conscious being. The narrative of "White Christmas" is philosophically important given its exploration of AI, personhood, and moral consideration which can guide our navigation of the future of AI as we tackle the technological advancements of AI and the accompanying moral dilemmas.

Concerning the episode "Be Right Back" (Brooker, 2013), the complexities of personhood, concerning the interactions between and treatment of AI and humans, are demonstrated in the narrative arc. In "Be Right Back", the protagonist, Martha, whose boyfriend Ash has passed away, makes use of a service for an AI clone of her boyfriend — first in the form of messages and later in the form of calls. She later purchases a physical android clone and uploads the AI into it. She begins to grow agitated with the clone for not being Ash: "You aren't you, are you? ... You're just a few ripples of you. There's no history to you. You're just a performance of stuff that he performed without thinking, and it's not enough" (*ibid.*). She is agitated by him following her orders, to the extent that she eventually tells him to jump off a cliff to kill himself. He starts to follow this order, but she is further upset that he is not begging for his life. Upon her request, he dutifully proceeds to beg her to spare his life, and she relents, stopping him from killing himself. Afterwards, she keeps him locked away in a closet, only allowed to come out and interact with her daughter on weekends. In this example, Martha views and treats robot-Ash as a mere tool to fill in for her deceased boyfriend, rather than an entity in its own right. If we view and treat the AI as a mere copy despite its behaviours and appearance equalling that of humans, we run the risk of unethical behaviour. If the AI is a perfect copy of a human in terms of appearance and is capable of movement in the world as another human, what is to distinguish AI from human? As previously discussed, a perfect AI copy is not ethically distinct from the original human. Here, however, the similarity is in appearance instead of behaviours as robot-Ash is physically indistinct from human-Ash but has different responses to situations from his human counterpart. As previously mentioned, even if we cannot be certain whether the AI possesses consciousness, if the behaviour is close enough, a practical ethical distinction is impossible. It ought to be morally impermissible to deny personhood and moral considerations on the basis of a distinction which is impossible to make. The ethics of such actions are brought into question in "Be Right Back" as this shows the negative consequences of classifying AI as "lesser than" and the consequences of treating AI accordingly. This science fiction example is philosophically interesting given it explores the consequences of treating AI as a mere copy instead of as an individual. Furthermore, it explores the moral questions surrounding the treatment of sentient beings when they are not awarded the same rights as humans. The narrative of "Be Right Back" is philosophically important given its exploration of AI, personhood, and moral consideration which we can use to guide our navigation of the future of AI as we tackle the technological advancements of AI and the resulting moral dilemmas thereof.

3.3. Why Science Fiction Should be Consulted as a Guide for AI, Personhood and Moral Consideration

These chosen examples of science fiction tackle the issues surrounding the personhood and moral considerations of AI within the likely near future and the technological singularity. While all three case studies display the negative impacts of denying AI personhood and moral consideration despite their autonomy, sapience, sentience, and perhaps even consciousness, some people are concerned that AI may be granted too much personhood and moral consideration. This alternate view is not often explored in science fiction and is found more in philosophical discussions pertaining to theory of mind and ethics. The concern at play in this view is that of "deception anxiety", i.e., the fear that AI will trick humans into thinking the AI is also human, thus deceiving the user (Dumouchel, 2022:2095). From this deception, the AI will have tricked the human into granting it too much personhood and moral consideration. In general, we are inclined to grant personhood and moral considerations to entities which look like us or are familiar in appearance, and which behave like us (particularly which display signs of suffering and happiness we find familiar). It is unlikely that this fear will be realised as it is uncommon for humans to mistake AI for other humans (*ibid.*).[4]

[3] I will not be touching on animal rights and the ethics thereof; I have mainly used it here as an analogy for the classification and treatment of AI in comparison to humans.

[4] Even in cases where individuals believe themselves to be in relationships with AI, they do not believe the AI to be human (Dumouchel, 2022:2095). This is evident in the fact that these individuals do not tend to introduce the AI to their friends and family, nor do they typically take the AI out on public dates.

The case studies discussed are of philosophical interest as they explore the possible future of AI, and are of philosophical importance as they offer insights on the morality of our decisions of whether to grant AI personhood and moral consideration. The issues highlighted here are of great importance to society as a whole moving forward in the upcoming years as AI becomes increasingly more sapient, more advanced, and more integrated in society and daily life. In the possible near future, we will be expected to make calls regarding AI in terms of their social, legal, and moral status. The science fiction narratives allow us to explore the implementation of certain principles and ideas by essentially observing them play out in society. This then allows us to weigh up the resulting consequences in ways standard philosophical investigation may not accommodate. It is for these reasons that such science fiction becomes topical and relevant to engage with for all members of society from policymakers to philosophers to everyday individuals. Hence, science fiction ought to be consulted as a guide for the future of AI concerning its personhood and moral consideration.

4. Conclusion

I have provided an introductory account of the connection of science fiction to moral philosophy. The discussion has demonstrated the value that science fiction has in particular for moral philosophy pertaining to issues surrounding AI, personhood, and moral consideration. Through the explorations of the science fiction narratives of the selected case studies, I have illustrated how science fiction can serve as a starting point for theorising about these issues. I have outlined how science fiction lends itself as an "enabling device" to explore and assess, question, and provide possible solutions to the philosophical issues of the future of society, AI ethics, the personhood of AI, and AI–human relations. Science fiction ought to be taken seriously, given its exploration of moral questions. Ultimately, my discussion has highlighted and stressed the importance of science fiction not only to philosophy but to society as a whole as it serves as a potential guide to navigating AI, personhood, and moral consideration in the possible futures.

References

Asimov, I. 1985. *Robots and Empire*. New York: Harper Voyager.

Asimov, I. 1995. *Gold: The Final Science Fiction Collection*. New York: Harper Voyager.

Asimov, I. 2000. *Bicentennial Man & Other Stories*. London: Orion.

Blackford, R. 2017. *Science Fiction and the Moral Imagination: Visions, Minds, Ethics*. Science and Fiction. Cham, Switzerland: Springer International. DOI: 10.1007/978-3-319-61685-8.

Brooker, C. (writ.). 2013. Be Right Back. *Black Mirror*, 2(1). 11 February. Channel 4.

Brooker, C. (writ.). 2014. White Christmas. *Black Mirror*, (7). 16 December. Channel 4.

Burton, J. 2015. *The Philosophy of Science Fiction: Henri Bergson and the Fabulations of Philip K. Dick*. London: Bloomsbury Academic.

Dumouchel, P. 2023. Ethics & Robotics, Embodiment and Vulnerability. *International Journal of Social Robotics*, 15(12):2087–2099. 1 December. DOI: 10.1007/s12369-022-00869-y.

Gomel, E. 2011. Science (Fiction) and Posthuman Ethics: Redefining the Human. *The European Legacy*, 16(3):339–354. 1 June. Routledge. DOI: 10.1080/10848770.2011.575597.

Mukerji, N. 2014. Why Moral Philosophers Should Watch Sci-Fi Movies, in Battaglia, F. & Weidenfeld, N. (eds.). *Roboethics in Film*. RoboLaw. Pisa: Pisa University Press. 79–92.

Tucker, A. 1996. Science Fiction as a Bridge to Philosophy (Review of How to Live Forever: Science Fiction and Philosophy by Stephen R. Clark). *Science Fiction Studies*, 23(3):534–536. 1 November. University of California Press. DOI: 10.1525/sfs.23.3.0534. https://www.jstor.org/stable/4240557.

Wolff, J. 2018. *An Introduction to Moral Philosophy*. New York: W.W. Norton & Company.