sciendo

Research Article

Reflecting on the Personality of Artificiality: Reading Asimov's Film *Bicentennial Man* through Machine Ethics

Jan Gresil S. Kahambing

Leyte Normal University
Tacloban City, Leyte, Philippines

Jabin J. Deguma

Cebu Technological University
Cebu City, Philippines

Doi: 10.2478/jesr-2019-0009

Abstract

The film Bicentennial Man (1999) pictured in a nutshell a robot who/that became human via his personality by plunging into the realities of freedom and death. The aim of this paper is to reflect on the notion of personality in the case of what this paper coins as a 'robot-incarnate' with the name Andrew, the first man who lived for two hundred years from his inception as an artificial machine. The method of exposition proceeds from (1) utilizing a philosophical reflection on the film concerning the determinacy of Andrew as a person and (2) then anchoring his case as a subject for the understanding of machine ethics. Regarding the first, the paper focuses on the questions of personality, death, and freedom. Regarding the second, the paper exposes the discussions of machine ethics and the issue of moral agency. Deducing from the already existing literature on the matter, the paper concludes that machine ethics must stand as the principle that serves as law and limitation to any scientific machine advancement showing promising potentials.

Keywords: *Bicentennial Man, Machine Ethics, Freedom, Personality, Artificiality, Asimov*

1. By Way of Reflection: Personality, Death, and Freedom

Asimov's *The Bicentennial Man and other stories* (1976) and later on expanded into *The Positronic Man* by himself & Silverberg (1992) backed the essential storyline of the film *Bicentennial Man* (1999). The film narrated the story of Andrew and the events that led him to becoming a human being. His first appearance as a robot, a machine coated in metal components from NDR series, happened when he (for the purpose of narrating, rather than an "it") was being welcomed in the Martin family as a helper in home maintenance. The father of the household, Sir Richard Martin, wanted Andrew to be recognized as more than just a tool in the family, so there were some adjustments of perception to be made before he was acknowledged as 'part' of the family. The deciding point of his cohabitation happened when, after breaking a figurine horse, a plaything of the youngest daughter Amanda whom Andrew addresses as 'Little Miss,' he started to carve from wood a similar figure of the horse as an apology to the young lady. The family then found out that, after inquiries in the NDR Company, Andrew is not the same as the other robots, having the ability to engage humor, sentimentality, and creativity. The first point of reflection here is his overall physiognomy and character: Andrew, apart from other robots, has something unique in him that makes him stand out from the rest – he has personality.

Here, one wonders whether the personality that Andrew possesses is a defect, a privation from other machines. This assumed defect, which in Andrew's case involves freedom, creativity, and sentimentality, what can best characterize this if not a random, out of the usual, formation that just happens in the make-up of a machine? The starting point of the reflection glares and revolves on this crucial genesis of a foreign substance: where does his personality come from? Is it built-in, or is his uniqueness, the character that composes his rare integral appearance, born of an unprecedented factory defect?

In this paper, Andrew's personality can be reflected on personology studies, focusing on personality as a whole through its 'components, levels, and spheres' (cf. Murray, 1938; Strack, 2005). To make a specification for its study, Warren's take on personhood can best serve the qualities of personality. Strictly, personhood is what makes a person a person, and personality is what makes a person unique as he/she is. This distinction can be blurred in the case of Andrew. In reflecting thus the personality that Andrew possesses, it is also an appeal at the same time on his case as a person.

Warren (1973) says that the central traits that best correspond to personhood can be roughly enumerated as: consciousness, reasoning, self-motivated activity, capacity for communication, and self-awareness. In order not to cause an overlapping of meaning, she further clarifies the terms. By consciousness, she means the capacity to feel pain. By reasoning, she means the capacity to solve new and 'relatively complex problems.' By self-motivated activity, she means the capacity to do activities that are relatively independent of 'either genetic or direct external control.' By communication, she means any form of means to relay messages of various types. By self-awareness, she means the presence of self-concepts.

Using Warren, Andrew, when he was still a robot, shows not only potentials but also vestiges of the traits. In the case of consciousness or the capacity to feel pain, Andrew, in an instance propelled by the broken toy of little miss, seems to take it upon himself to make amends, showing signs of sensitivity and emotion, pained as it were at the relationship that he had had with the sweet and caring youngest daughter of the Martin family. Later on in the film, Andrew, externally appearing now as human but with prosthetic machine organs, felt something that he immediately recognized after Rupert Burns, son of the robot designer, made him aware of the feeling of emotional pain at the sight of one's beloved in the arms of another – jealousy. In the case of reasoning, Andrew surely showed signs of problem-solving skills at repairing the old gramophone at the basement. As the film went on, he simply resolves every problem that comes his way, including in the end his ultimate problem of being accepted as a human being. In the case of self-motivated activity, this would be the mark that hits Andrew's personality: apart from other robots, his actions are clearly not from what he was first made of, not out of genetic or manufactured composition in the case of robots. His signs of creativity and growth in knowledge is indicative of the personality that self-motivates to become who he wants to be. And he relates this through communication. At first, he had a hard time dealing with jokes – as jokes without an author seem to require a higher level of social understanding and context. But he did it, and he understood humor as a form of communication. And finally, in the case of self-awareness, Andrew is well aware of himself, what he needs and what he must do. He even transcends what is beyond what a robot can do – to love.

The first point of reflection then on the randomness of Andrew's personality seems to be connected to the human traits of personality. And this puts a question mark on Boethius' idea of a person as an individual substance with a rational nature. Seeing Andrew's personhood as a robot, it begs the question for humans, if seen in reverse for robots: what if humans are also accidents and not substances?

The second point of reflection is the existential notion of freedom for Andrew. The events that led to this is first his naivety that little miss actually has a certain affection for him but she cannot, given that Andrew is only a robot, a machine. And this is the next step at properly identifying the locus of himself. In the film, Andrew does refer to himself as 'it' or 'I' – this I, the subjectivity that is assumed to indicate the radical wholeness of the self as itself, is missing in Andrew's orientation as a robot. Instead, he refers to himself as 'One': "One is glad to be of service," "One understands why some animals eat their young," "One has studied your history. Terrible wars have been fought where millions have died for one idea, freedom. And it seems that something that means so much to so many people would be worth having." The quest for this identity within his personality however is answered as the film went on: his showing of signs of intentionality in his first expression of 'I'

becomes explicit in his want to buy his freedom. As a precondition then to love, the human condition is ravaging, if not clamoring, for the freedom to express what it feels. It is in this setup that the existential notion of love must be situated: it must first rest within the realm of one's freedom. Incidentally however, the manner in which he succeeded in achieving it is through the ideology that dictated his time, in the operations of capitalism, that is, money bought for him his freedom. Even a little child in 'little miss' can express that Andrew needs his own bank account, something odd for a machine to have, to connote that in the world of capitalism, creativity and the products that come from it must be monetarily compensated; everything as it were can be opportunities for private profit. The total package of Andrew's freedom is given to him by Richard Martin but as one lesson in psychoanalysis goes, freedom comes with a price, so Andrew is banished from the Martin household. After he goes out to explore more his identity, by building a house and living on his own, he embraced change. The foremost of this change is the death of his master. After he realized this, he soon reflects that death is going to be the constant finality of existence for humanity.

The third point of reflection can be centered on the question of death. Quite the anti-thesis of immortality, Andrew bargains his being an immortal entity by becoming human. He perfects this first by adding prosthetic organs in his body, an artificial liver, kidney, and even stomach. The fundamental element of nutritive activity remains as a crucial step in becoming human. The so-called threshold of survival in food, drink, is completed in sex. Here, the paradigmatic shift that stands at the climax of Andrew's humanity is not the epidermis appearance of Andrew, but his capabilities of attaining the basic survival instincts of a human person: eating, drinking, and sexual intercourse. The last element was completed in his affair with Portia, the granddaughter of little miss. Herein lies the extent that reaches the query on human oblivion and finitude: the question of death is not separated from the question of love: One can never love a perfect person, and to be human is to commit mistakes, to do the wrong thing. In Andrew's words,

> *That you can lose yourself. Everything. All boundaries. All time. That two bodies can become so mixed up, that you don't know who's who or what's what. And just when the sweet confusion is so intense you think you're gonna die... you kind of do. Leaving you alone in your separate body, but the one you love is still there. That's a miracle. You can go to heaven and come back alive. You can go back anytime you want with the one you love.*

Out of his love for Portia and the recognition of being human, he seems to make the act of sacrificing an eternal existence: this time not in the case of an angel becoming man as in *The City of Angels,* but a robot becoming man. He accomplishes this in establishing a neural system and circulating blood to make him age. Seen in the perspective of humanity, and in fact in the first Congress that judges on the case of Andrew, humanity cannot accept Andrew as human mainly because he cannot die. An immortal existence is something enviable for a human, and for Andrew to claim his acknowledgment, he has to pay the price of his life. To pay such price is for a human a mistake, since immortality is a goal that humans have for a long time dreamed of attaining. In the process of attaining death, Andrew fully assumed the term 'robot-incarnate'- a robot made flesh.

But that is the final point of reflection that this paper exposes. What goes beyond death seems to be an odd topic for a machine. Questions of eschatology like 'will Andrew go to heaven or hell?' surely is another story. In contemporary society however, machines are now playing a major role. They are a special area of interest since they cross and erode boundaries that have never been crossed before except for lofty imaginations in science fictions. The mechanism that one utilizes in robots is unique in a sense that it applies to almost any technological aspect of society's growing dependence of it. And here, one glares intently for example at the fact that a robot can worship, which surely is a sight that eradicates the very notion of spirituality in a human subject.

For instance, the majority of Buddhist priests in Japan who take part-time jobs owing to the declining interest in religion triggered the filling in of a humanoid robot called "Pepper" that can read sutras or Buddhist scriptures at funeral ceremonies (Thuy Ong, 2017). Compared to hiring a priest, Pepper costs 20% lesser. Later, SoftBank Group Corp. upgraded Pepper to take orders in English and Chinese to be offered to restaurants and other businesses (Jiji, 2017). What this first brings into attention is the question of limitation within technological advancements: where does the border lie? Here is where machine ethics comes in.

2. Machine Ethics: Laws and Limits

The advent of the AI or artificial intelligence, one in which Andrew's personality may possibly reside, can be contextualized in Asimov's "Three Laws of Robotics," namely:

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.

Although seemingly basic and comprehensive, this ideal set of rules for such machines is "an unsatisfactory basis for Machine Ethics, regardless of the status of the machine (Anderson, 2016)." Critics comment on this set of laws, arguing that if a real robotic existence with personality might soon emerge in the future, one will have to thoroughly go over such laws. Robots are becoming vital in the future economy (International Federation of Robotics, IFR, 2016), owing to some description of the present age as the *age of robots* (Veruggio, G. et.al, 2016). To describe this landscape is to acknowledge that the time of machines has not only dawned but is already at its noon time. Without consensus and proper distinction, scholars initially coined the emergence of such terms as *roboethics* which "deals with the issues derived from the diverse applications of robots understood as particular devices" and of *machine ethics*, which "studies how to endow self-regulated machines with ethical behavior." The term device here is ambiguous inasmuch as it is vital, such that an *autonomous device* can be used interchangeably with robot, machine, autonomous machine, or automaton (Lumbreras, 2017).

In another study, Allué (2003) distinguishes what is a machine, robot, and cyborg. The general characteristic of a machine is a mechanism, a tool. A robot is the borderline position between a man and machine, having organs for efficiency like arms and legs in order to perform tasks and even 'central decision organs' analogous to a human's memory. "The robot was a machine capable of imitating human beings; the cyborg was the actual combination of machine and human (Allué, 2003, 21)." She then expounds that Andrew's case is a transition from being a machine, a tool, into becoming a robot when Andrew got his prosthetic organs, and then into a cyborg, a combination of human and machine with his neural and circulatory system. But this combination as it were produces a lot of complications. Even the relationship between Andrew and Portia seems an unlikely relation. Allué concludes that Andrew as a cyborg acknowledged as a human still calls for the limitation of machine ethics.

The limitation of machine ethics lies in the programming language of a machine: while it can have objectives or ends reflective of the ends of man, it cannot have its own subjective experience. The emergence of values, although mired by possibilities, is a mystified manner of optimistically dealing with the insistence that a machine can have its own – unique, personal, and conscious – character. That is to say, that if there is a limitation to suppose the subjective experience of a machine to refer to values and sentimentality, it is that what keeps it from attaining such is its very own programming language distinctly separate from human personal language. Along the lines of machine ethics therefore, the very notion of personality is not going to be a realistic mark to make-up the obvious individuality of a machine – that it can have its own freedom and creativity – since at best, it can only operate on the language that the programmer arrays to it, coded it may be in a lengthy series of formulations or bulky manuals that are constantly updated.

For Lumbreras (2017, 8), "nothing suggests that automata based on current technology will ever be able to deal with personal language (which seems to be linked to subjectivity) or have a subjective experience of their purpose—at the very least, there is nothing to make us think otherwise." This conclusion seems plausible except that in the emergence of Sophia – the social humanoid robot declared as a Saudi Arabian citizen (*UNDP in Asia and the Pacific*, 2017) – such position ought to be reevaluated. But her responses and the language that constitutes her understanding is still suspect. Critics slam the presentation of Sophia as having consciousness, which is, appearing only to show signs of human intelligence. Gunkel (2012) emphasizes on the significance of questioning, dedicating his book on the critical perspectives of asking the right

questions for machines. Can machines really engage in a truly ethical embodiment, capable of consciously evolving into agents of personal and social transformation? If there are two sides in the spectrum to evaluate Sophia's (and perhaps too, Andrew's) authenticity, they would have to be the optimistic side and the negative side.

On a positive note, Kurzweil (2012) believes that authenticity born out of the consciousness of machines will emerge in their experiencing of values, which then can be branded as *spiritual machines.* What Kurzweil seems to suggest in the extreme is that there can be an emergence of spirituality in machines: that this is a possibility that optimistically will happen. He says

> *Regardless of the nature and derivation of a mental experience, spiritual or otherwise, once we have access to the computational processes that give rise to it, we have the opportunity to understand its neurological correlates. With the understanding of our mental processes will come the opportunity to capture our intellectual, emotional, and spiritual experiences, to call them up at will, and to enhance them (Kurzweil, 1999, 109; Italics mine).*

Kurzweil is then positive of the fact that there is a substratum in which one can fully analyze consciousness to the point of imposing it as blueprint for any future evolution in the technological advancement of personality in machines.

On the other side however, those critical of the agency of a machine as having a concept of morality argue again of the limitation of the programming language. Concerning the dimensions of learning and morality in machines, there is no wide-ranged calculation whatsoever in the current situation to have come up the notion of moral machines acting as ethical agents that can rationally justify, argue or even make the point of deductive conclusions in some instances why certain actions must be taken.

> *Morality is to be applied in a wide scope of situations. Big-data analysis consists of searching for buried patterns that have some kind of predictive power. However, one can't base machine ethics on predicting how to do the right thing. Morality is not about statistics, but about being right or wrong according to rules and exceptions, in specific circumstances. Present day wide scope machine learning from big data does not come up with rules, and so cannot explain and justify actions, which is a must for ethical machine actions to be accepted. The community should be well aware that such present day learning is inadequate for general machine morality. Only small, circumscribed, well-defined domains have been susceptible to rule generation through machine learning. Rules are all important for moral explanation, justification and argumentation (Pereira & Saptawijaya, 2016, 170-171).*

The booming enterprise of sex robots, for instance, catered the way for a sensitive area in machine ethics. Bendel (2017, 25) carefully notes that machine ethics can be helpful in thoroughly going over the questions that concern in the construction of sex robots, "which are moral machines in their appearance and in their behavior." Critics abhor the sudden growth in the production of such robots, arguing instead for the utilization of funds for medical purposes and other activities to promote the common good in social welfare states. However, even when the focus will be shifted primarily on the production of an AI in the domain of machine ethics, it still poses a dilemma.

The glaring problem of an AI capable of full integration to strict ethical standards will induce the mythical conceptions of a machine killing its own creator like Frankenstein. "If an AI possessed any one of these skills—social abilities, technological development, economic ability—at a superhuman level, it is quite likely that it would quickly come to dominate our world in one way or another (Armstrong, 2014, 15)."

In such a scenario, a strong AI poses a danger that surpasses the options of whether to teach it ethics or not: it exposes the inevitable end of humanity either when robots enforce ethics or when it has its own agency. Because humans constantly violate the laws of ethics which they themselves promote, the AI's black-and-white moral standards will surely lead to the demise of humans who are supposed to be practitioners of virtue, always falling short of the norms the AI will come to understand as it evolves. If this is pursued, the imposition of robots to society will have immense consequences similar to the totalitarian regimes of modernity.

The reality that resists in this conception is the idea that ethics will still be the principle of limit.

To ensure safety, "an AI will likely need to be given an extremely precise and complete definition of proper behavior, but it is very hard to do so (Armstrong, 2014, 43)." Lumbreras proposes filtered decision making as a possible solution (2017, 4). She basically says that this filtering of decisions consists in designing the learning-based automaton in such a way that it can obtain inputs from the external world and process them to get a decision, but not activate that decision by itself. Instead of determining the output of the process directly, the learning automaton would send the preliminary output to an external, objective module, which would be accessed online. That is to say, humanity will still decide on the decision of the AI before it can execute the decision itself.

What shall be the basis for this decision? Edgar (2003) situates his understanding of machine ethics in the Four Causes of Aristotle. For him, it will be helpful in scientific research to find out the purpose of change. He first enumerates the causes: "(1) the *material* which is changed, (2) the *efficient* cause of the change (presumably the scientist, or some natural force the scientist sets in motion), (3) the *formal* cause of the change (what was the plan, or blueprint), and (4) the *final* cause of the change (its purpose, or goal)." He argues that among the causes, scientists devote more time on the material cause and the efficient cause, and then coming up with the formal cause, but there is a lack in the attention to the final cause. The question then "should not be just '*How* can we do this?' but more importantly, '*Why* should we do this? (Edgar, 2003, 390)."

It is important therefore to understand the rationale of making machines into moral agents rather than finding out how. If this were a matter of 'how,' there would be conflicting and overlapping viewpoints. In a *Top-Down* ethics for machines, one is led to the consequence of procuring a large amount of data to program a machine replete with the variables that will guide decisions in ethical dilemmas. But to expect for this 'pure philosophical investigation' is overly optimistic. In a *Bottom-Up* ethical approach, the machine evolves according to the moral norms it sees in society. Bottom-up ethics is "where the ethical concepts are defined using the working processes and architecture of the machine (Castro, 2016)." However, this would involve jeopardizing the situation when the machine-agent might understand the wrong values, if seen only from the set of moral contexts with harmful results (cf. Yudkowsky 2008)." A good example of this is Chapie, the robot from the film of its namesake, when it evolved in the hands of robbers and bandits. Another would be Norman 'the psychopath AI' (Cuthbertson, 2018). Shuman, et.al, then suggests

> *that an external boost to direct ethical theorizing is required. Efforts in the fields of neuroscience, experimental philosophy, and moral psychology have recently provided powerful insights into the structure of our moral values and intuitions (Haidt and Graham 2007; Koenigs et al. 2007; Knobe 2003), and it is reasonable to expect further gains. (Shulman, 2009, 'Which Consequentialism')*

Yet again, the limitation of this rationale falls flat when it insists on absolute autonomy, which even humans do not fully have with all the perennial problems of free will and choice. For Sullins (2011, 153), "no robot in the real world – or that of the near future – is, or will be, as cognitively robust as a guide dog. Yet even at the modest capabilities of today's robots, some have more in common with the guide dog than with a simple tool like a hammer." In robotics technology, the schematic for the moral relationship between the agents is:

Programmer(s) → Robot → User

This setup again dictates that there is no freedom for a machine, and that its choices will still come from the programmer who codes its language. Even in the writers of science fiction and those who conceive of robots into becoming humans, the cyborg as metaphor for future humanity, there are serious consequences that need to be addressed, dangers that might prove to be blindsides of the situation. This is crucial "especially in an age in which language itself seemed to be complicit in blurring the distinction between man and machine, to the fraught potential of language, and to the consequences of those dangerous potentials to their own writing (Cook, 2007, 136)."

The danger that lies in the corner of an autonomous machine is something that might be uncalled for. It is in this light that Asimov composed his 1985 Revised Laws of Robotics. He inscribes a precedent law before all other laws: which he calls 'The Zeroth Law': A robot may not injure humanity, or, through inaction, allow humanity to come to harm.

> *After introducing the original three laws, Asimov detected, as early as 1950, a need to extend the*

> *First Law, which protected individual humans, so that it would protect humanity as a whole. Thus, his calculating machines "have the good of humanity at heart through the overwhelming force of the First Law of Robotics." In 1985, he developed this idea further by postulating a "zeroth" law that placed humanity's interests above those of any individual while retaining a high value on individual human life. (Clarke, 2011, 268ff)*

The endpoint of machine ethics' laws and limitation then lies in the idea that the robot must not only injure a human being, but humanity as a whole. Accordingly, such ethics must look at humanity not as an antinomy for machines, but as its creator. The scenario that this calls for is a careful look at the consequences of this invention in the future. It would be exhaustive, but it is a limit that can save humanity itself.

> *Since such development may require extensive research and it is not currently known when such procedures will be needed to guide the construction of very powerful agents, the field of machine ethics should begin to investigate the topic in greater depth (Shulman, et.al. Machine Ethics and Superintelligence, 97).*

Ethics as a principle of limit should inform, as well as guide, the fast-paced growth of scientific advancements in technology. That is to say, "machine ethics research may have some social value, but it should be analysed in a broader lens of the inherent difficulty of intelligent action in general and the complex social context in which humans and computational agents will find themselves in the future (Brundage, 2014)." There would be no problem therefore if, in the future, humanity can find Andrews who share human personalities, the inherent goodness that he possesses and that might even surpass the goodness of some humans who disturb societal norms for selfish reasons. But this future is as complex as it seems, and humanity must prepare for it in the unlikely event that one Andrew may turn out to be an evil genius presaging the apocalypse of human oblivion.

3. Conclusion

The paper exposed the film *Bicentennial Man* (1999) by picturing the realities of Andrew and his personality. This is done first by utilizing a philosophical reflection with three points on the film concerning the determinacy of Andrew as a person, the question of freedom, and then of death. Next, it tackled the case in the perspective of machine ethics. Machine ethics serves as a framework in which cases of the blurring distinctions of man and machine and the machines for the future can be captured. Machine ethics must stand as the principle that serves as law and limitation to any scientific advancement showing dangerous potentials. A machine in the growing progress of science needs an ethical component that must serve as laws to limit the possible adverse effects of its production and eventual global use in the future.

References

Allué, S. (2003). Blurring Posthuman Identities: the New Version of Humanity Offered by Bicentennial Man (1999). *Odisea 4*, 17-30.

Anderson, S.L. (2016) Asimov's "Three Laws of Robotics" and Machine Metaethics. *Science Fiction and Philosophy: From Time Travel to Superintelligence, second edition.* Wiley.

Armstrong, S. (2014). *Smarter Than Us: The Rise of Machine Intelligence.* USA: Machine Intelligence Research Institute.

Asimov, I. (1976). *The Bicentennial Man and Other Stories.* US: Double day.

Asimov, I., & Silverberg, R. (1992). *The Positronic Man.* UK: Gollancz.

Bendel, O. (2017). Sex Robots from the Perspective of Machine Ethics. In *Love and Sex with Robots. Second International Conference, LSR 2016, London, UK,* Revised Selected Papers (Cheok, A.D.; Devlin, K.; Levy, D. (Eds.), 17-26.

Brundage, M. (2014). Limitations and Risks of Machine Ethics. In *Journal of Experimental & Theoretical Artificial Intelligence 26*(3).

Castro, J. (2016). A Bottom-Up Approach to Machine Ethics. DOI: http://dx.doi.org/10.7551/ 978-0-262-33936-0-ch113

Clarke, R. (2011). Asimov's Laws of Robotics: Implications for Information Technology. In In *Machine Ethics.* (Anderson, M., Anderson, S.L. Eds.). Cambridge: Cambridge University Press.

Cook, J.C. (2007). *Machine and Metaphor: The Ethics of Language in American Realism* (Cain, W. ed.). In *Literary Criticism and Cultural Theory*. Routledge.

Cuthbertson, A. (2018). Meet Norman, The 'Psychopath AI' That's Here to Teach us a Lesson. *Independent.* Retrieved from https://www.independent.co.uk/life-style/gadgets-and-tech/news/norman-psychopath-ai-bias-mit-artificial-intelligence-reddit-a8389011.html/amp.

Edgar, S. (2003). *Morality and Machines: Perspectives on Computer Ethics.* Second Edition. Sudbury, Massachusetts: Jones and Bartlett Publishers.

Gunkel, D. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics.* London, England: The MIT Press.

Haidt, J. and Graham. J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize. *Social Justice Research 20*(1), 98–116.

International Federation of Robotics, IFR, (2016). World Robotics 2016. Frankfurt: International Federation of Robotics.

Jiji (2017). Softbank upgrades humanoid robot Peper. *The Japan Times.* Retrieved from https://www.japantimes.co.jp/news/2017/11/21/business/tech/softbank-upgrades-humanoid-robot-pepper/#.Wm51Vq6nHIU

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis 63*(3), 190– 194.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., and Damasio. A. (2007). Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements. *Nature 446*(7138), 908–911.

Kurzweil, R. (1999). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence.* England: Viking Penguin.

Kurzweil, R. (2012). *How to Create a Mind: The Secret of Human Thought Revealed*. London: Penguin.

Lumbreras, S. (2017). The Limits of Machine Ethics. *Religions 8*(100), 1-10.

Murray, H.A. (1938). *Explorations in Personality*. New York: Oxford University Press.

Ong, T. (2017). Pepper the robot is now a Buddhist priest programmed to chant at Funerals. *The Verge*. Retrieved from https://www.theverge.com/2017/8/24/16196752/robot-buddhist-priest-funeral-softbank

Pereira, L. M., Saptawijaya, A. (2016). Programme Machine Ethics. In *Studies in Applied Philosophy, Epistemology, and Rational Ethics* vol. 26. Switzerland: Springer International Publishing.

Shulman, C., H. Jonsson, N. Tarleton (2009). Machine Ethics and Superintelligence. In *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan, Proceedings,* edited by Carson Reynolds and Alvaro Cassinelli, 95–97.

Shulman, C., N. Tarleton, H. Jonsson (2009). Which Consequentialism? Machine Ethics and Moral Divergence. In *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan, Proceedings*, edited by Carson Reynolds and Alvaro Cassinelli, 23–25.

Strack, S. (2005). *Handbook of Personology and Psychopathology*. Wiley

Sullins, J. (2011). When is a Robot a Moral Agent? In *Machine Ethics.* (Anderson, M., Anderson, S.L. Eds.). Cambridge: Cambridge University Press.

Veruggio, G, Operto, F., and Bekey, G. (2016). Roboethics: Social and ethical implications. In *Springer Handbook of Robotics*. Berlin and Heidelberg: Springer, pp. 2135–60.

Warren, M. A. (1973). On the Moral and Legal Status of Abortion. *The Monist 57*(1).

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, (pp. 308–345). New York: Oxford University Press.