

# Globethics Repository

The logo for Globethics, featuring the word "Globethics" in white, sans-serif font centered within a solid blue rectangular background.

## Confucian robotic ethics

This page was generated automatically upon download from the Globethics Repository. More information on Globethics see <https://www.globethics.net>. Data and content policy of Globethics Repository see <https://repository.globethics.net/pages/policy>.

Item Type	Book chapter
Authors	Liu, JeeLoo
DOI	<a href="https://doi.org/10.58863/20.500.12424/4276019">10.58863/20.500.12424/4276019</a>
Publisher	Globethics Publications
Rights	Globethics Publications;Attribution-NonCommercial-NoDerivatives 4.0 International
Download date	2026-04-16 17:57:42
Item License	<a href="http://creativecommons.org/licenses/by-nc-nd/4.0/">http://creativecommons.org/licenses/by-nc-nd/4.0/</a>
Link to Item	<a href="http://hdl.handle.net/20.500.12424/4276019">http://hdl.handle.net/20.500.12424/4276019</a>

## CONFUCIAN ROBOTIC ETHICS

*JeeLoo Liu, USA*<sup>221</sup>

### **Abstract**

This paper will explore the possibility of implementing Confucian ethical codes into the so-called AMAs (artificial moral agents). Drawing from the Confucian classic *The Analects*, it will consider what ethical precepts could be incorporated into robot morality. It will also contrast Kantian AMAs, Utilitarian AMAs, and Confucian AMAs to decide the strength and weakness of each model. The paper's thesis is that robots should be constructed with certain virtues highlighted in Confucian virtue ethics. With the Confucian moral codes built in, their functional ethics can qualify them as moral agents, albeit artificial.

---

<sup>221</sup> JeeLoo Liu, Department of Philosophy, California State University, Fullerton/USA. Jeelooliu2gmail.com. © Globethics Publications, 2023 | DOI: 10.58863/20.500.12424/4276019 | CC BY-NC-ND 4.0 International.

## 9.1 Introduction

With the advancement of AI technology, the appearance of intelligent humanoid robots in our society is very likely in the foreseeable future. Whether they truly possess human intelligence and can think like humans do is up to philosophical debate, but they will surely pass the Turing Test—i.e., they will entice their human interlocutor to be inclined to treat them as humans. Intelligent robots will one day become members of our society, sharing our jobs, taking care of our elderly, serving us at restaurants and hotels, making important navigational, military and even medical decisions for us. Should we equip these robots with a moral code to teach them right from wrong? If so, what kind of moral codes would be able to bring about the kind of artificial moral agents (AMAs) that we would like to have in our society?

On the optimistic assumption shared by many AI designers that the development of AMAs can be successful one day, this paper will explore the possibility of implementing Confucian ethical codes into the so-called AMAs. Drawing from the Confucian classic *The Analects*, it will consider what ethical precepts could be incorporated into robot morality. It will also contrast Kantian AMAs, Utilitarian AMAs, and Confucian AMAs to decide the strength and weakness of each model. The paper's thesis is that even though robots cannot have our innate moral sentiments, the four moral sprouts that Mencius defends, they can be constructed with the kind of ethical principles that Confucianism stresses. With the Confucian moral codes built in, their functional ethics can qualify them as moral agents, albeit artificial.

The investigation of AI ethical codes is not just a futuristic mind-game. According to Michael Anderson and Susan Leigh Anderson, "Machine ethics, by making ethics more precise than it has ever been before, could lead to the discovery of problems with current ethical theories, advancing our thinking about ethics in general" (Anderson & Anderson 2006, 11). This paper will show that the comparative study on

robot morality can shed light on the flaws of the Kantian model as well as the Utilitarian model for human ethics.

## **9.2 The Rise of Machine Ethics**

Up to now, having robots that can make systematic ethical decisions with advance considerations of their consequences is still a remote dream. However, there are already existing specific decision guidelines for machines. Some of these decisions have morally significant consequences. For example, autonomous military drones can be programmed to strike or withhold attack with the detection of civilians in the vicinity of a military target. Health-care robots can also be programmed to take life-saving measures or to forego further treatments. According to Ryan Tonkens, “Because autonomous machines will perform ethically relevant actions, akin to humans, prudence dictates that we design them to act morally” (Tonkens 2009, 422). Therefore, even if we cannot make “moral machines” yet, we must consider machine ethics. Furthermore, the version of machine ethics we formulate should be applicable to foreseeable robotic moral reasoners, and not just to their programs designed by humans. In other words, machine ethics is concerned with applying ethical codes to artificial moral agents, not to their designers.

According to Allen, Smit and Wallach (2005), there are three fundamentally different approaches to designing artificial morality: the bottom-up approach, the top-down approach, and a hybrid model that combines the above two.<sup>222</sup> The bottom-up approach is to have the machine develop its own ethical code from the piecemeal rules in its day-to-day decisions. The machine can be given learning skills with which to process the information gathered by facing the consequences of various courses of action it takes. To promote a certain type of behaviour, the designer can create a reward system that favours certain actions that the

---

<sup>222</sup> There are of course hybrid approaches as well.

machine takes. Such feedbacks can enable the machine to develop its own ethical codes in time. This approach is similar to human childhood learning experience in building moral character. However, Allen et al argue that it is questionable whether this approach could be helpful in developing artificial moral agents that are “capable of engaging the more complex dilemmas that we encounter daily” (Allen et al 2005, 152). The top-down approach, on the other hand, is to implement general, abstract ethical rules that would govern the machine’s daily decisions and actions. To use this approach, the designer must first choose an ethical theory to analyse “the informational and procedural requirements necessary to implement this theory in a computer system,” and then design its subsystems for the implementation of the ethical theory (Wallach & Allen 2009, 80). Even with the preset design, in each scenario the machine will need to use deduction to determine the best course of action under the programmed ethical principles. This approach reflects the debates in normative ethics, since different ethical theories will generate different ethical codes for artificial moral agents. In this paper, we will first consider three leading models: Asimov’s Laws of Robotics, Kantian deontology, and utilitarianism. To fully evaluate the applicability or the desirability of each ethical model, Allen et al suggest that we must give “careful consideration to the prospects for building AMAs by implementing decision procedures that are modelled on explicit moral theories” (Allen et al 2005, 150). In other words, the devil is in the details. However, this paper will not be able to touch on the technical implementation of ethical principles or the algorithmic design of the decision procedure. The critique will be largely conceptual.

In this paper, we suggest a hybrid approach, combining both some general ethical principles in the robot’s initial design, and a learning mechanism that enables the robot to improve and improvise as it goes through different trial and actual situations. According to Anderson & Anderson (2007), the aim of machine ethics is to define explicit ethical principles that artificial intelligence could appeal to in choosing and

justifying its own actions. They argue that we cannot possibly give specific rules for each and every possible situation that might come up. “The virtue of having principles to follow, rather than being programmed in an ad hoc fashion to behave correctly in specific situations, is that it allows machines to have a way to determine the ethically correct action in new situations, even in new domains.” (Anderson & Anderson 2007, 17) In other words, we want to have artificial intelligence to actually be artificial moral agents with their own moral principles, making moral deliberations on the basis of those principles, and justifying their action by appealing to those principles. Hence, the choice of a set of moral principles that can be implemented into artificial intelligence is a crucial task of machine ethics. Having only the abstract ethical principles is not sufficient to equip machines with the abilities to adapt to new situations. Virtue ethics defines ethical behaviour as what a virtuous person would perform in the given situation. “A virtuous person is defined as a person who has learned and internalised a set of habits or traits termed virtuous. For a virtuous person, virtuous acts become second-nature, and hence are performed in many different situations, through time” (Govindarajulu et al 2019, 1).

The advantage of virtue ethics to the deontological and the utilitarian approaches is that virtue ethics can better capture human responses, human expectations, and human values. Virtues are the stable, consistent traits that can have explanatory power for the agent’s behaviour and can be used with predictive power for the agent’s future behaviour (Alfano 2013). Govindarajulu et al argue that “if the conditions of stability, consistency, explanatory power, and predictive power hold, then virtuous agents or robots might be easier for humans to understand and interact with (compared to consequentialist or deontological agents or robots)” (Govindarajulu et al 2019, 2). That is to say, virtue ethics is more conducive to the design of sociable robots, the kind of robots that can interact and communicate with humans. Virtue ethics provides a hybrid model in that the machine is designed with certain characteristic

traits (virtues) that enable it to learn from the available data, and to make novel decisions consistent with the preset traits. Instead of giving robots abstract ethical principles to add into its processing and calculation to generate action, the approach of virtue ethics equips robots with some predetermined traits (their “virtues”), and a large databank that stores information of what virtuous agents have done or would do. Before we begin to collect data on virtuous agents’ behaviour, what we need to identify first is which virtues are essential to sociable robots.

This paper will introduce Confucian virtue ethics, and lay out some essential virtues that artificial moral agents should have in order to be accepted into an ethical human society.

### **9.3 The Trolley Problem and Various Ethical Models**

Since intentional robot action or intervention is still a hypothetical scenario, we might find it helpful to appeal to the commonly used thought experiment for ethical dilemma: the trolley problem.

#### **9.3.1 The Standard Trolley Problem**

A runaway trolley is rushing down the railway tracks, and there are five people on the tracks ahead, unable to flee in time. The robot safety inspector (or driver) can intervene by pulling the lever to divert the trolley onto a different track. However, there is also one person on the other track. The choice is between sacrificing one person to save five human lives, and not sacrificing one life and letting the five people die. Should the robot inspector/driver pull the lever to prevent the disaster, or do nothing?

#### **9.3.2 The Footbridge Variation of the Trolley Problem**

The robot safety inspector is standing on the footbridge above the trolley track to observe the trolley traffic. Upon seeing that under the bridge a runaway trolley is heading down a track with five people stuck

on the track, the robot has to do something quick. Next to it on the footbridge is a heavyset man observing the same scene. If the robot pushed the man onto the track to stop the trolley, it would be able to prevent the disaster of having five people killed. Should it do it?

Faced with the two kinds of dilemma, the moral agent must decide “whether the action required to save the five is impermissible because it causes harm, or permissible because the harm is only a side effect of causing good” (Deng 2016). Experiments show that humans typically choose to sacrifice one to save the five people in the trolley example, but would not choose to sacrifice the fat man on the bridge to save the five people on the track. According to the experimenters: “This leaves psychologists with a puzzle of their own: How is that nearly everyone manages to conclude that it is acceptable to sacrifice one life for five in the trolley dilemma but not in the footbridge dilemma, in spite of the fact that a satisfying justification for distinguishing between these two cases is remarkably difficult to find” (Greene et al 2001, 2106)? These two cases will serve as our test case for the various ethical models for artificial intelligence.

In human ethical contexts, the trolley problem may seem far-fetched and unrealistic; however, in the context of AMAs, similar situations may arise with machine ethics. Imagine a future Tesla equipped with an ethical overriding principle to avoid harming more people than necessary.<sup>223</sup> Suppose a school bus filled with school children suddenly loses control and is crashing into a Tesla, which cannot stop in time to prevent a collision. Should the car veer off to hit a median, risking the life of its

---

<sup>223</sup> Examples such as this are abundant in the discussion on self-driving or driverless cars. See Bonnefon *et al.* 2016, Deng 2015, Greenemeier 2016, and Herkewitz 2016. Of course, currently driverless cars do not make ethical decisions on their own. They can only take in information about speed, road conditions, weather conditions, and so on, to make instantaneous driving decisions. But if artificial moral agents are possible, then some simplified ethical codes could conceivably be programed into these cars as well.

driver, or should it continue its course and let the collision take place? Maybe no one would want to buy a Tesla if given this kind of moral consideration, it could end up sacrificing its driver, but the point is that this scenario is analogous to the trolley problem cases. Many such scenarios could be envisioned. Therefore, the trolley problem could serve as the test for our ethical theories.

### *9.3.2.1 Asimov's Laws of Robotics*

A prime example of the top-down approach is to appeal to the Three Laws of Robotics composed by Isaac Asimov in 1942<sup>224</sup>:

*[A1] A robot may not injure a human being or, through inaction, allow a human being to come to harm.*

*[A2] A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.*

*[A3] A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. (Cited in Wallach & Allen 2009, 34)*

A fourth law was later added, which supersedes the above three:

*[A4] A robot may not harm humanity, or, by inaction, allow humanity to come to harm. (Wallach & Allen 2009, 91)*

The difference between the First Law and the Zeroth Law is that the former concerns individual human beings while the latter concerns the general humanity. Since the Zeroth Law supersedes the First Law, the implication is that the robot may conceivably harm individual human beings if doing so could prevent harm to humanity. An example of the application of this law in a science fictional situation would be if some individuals carry deadly contagious virus that could potentially wipe out humankind, then the robot would be obligated to eliminate such individ-

---

<sup>224</sup> The Three Laws were first introduced in Asimov's science fiction story *Runaround* in 1942. According to Wallach & Allen (2009), "No discussion of top-down morality for robots can ignore Asimov's Three Laws." (Wallach & Allen 2009, 91)

ual humans. However, such a moral precept is highly dubious, since the notion of humanity is rather abstract and has been used to justify many evil practices in human history such as eugenics, ethnic cleansing, and so on. Having an overriding Zeroth Law thus can conceivably annul Asimov's Three Laws.

When applied to the trolley problem, Asimov's Three Laws are clearly inadequate. In the standard trolley scenario, the First Law prohibits the robot's pulling the lever, as it will bring harm to the one person on the other track, but it will also prohibit the robot's doing nothing, as its inaction will lead to the harm of five human beings. In the footbridge scenario, the robot's pushing the fat man over the bridge is strictly prohibited, since it directly involves harming a human being; but at the same time, the robot's doing nothing, when the option of pushing the fat man over to stop the trolley is available, is also against the second clause of the First Law. Either way, the robot is in a bind with no ethical guidance. Alan Winfield et al conducted a series of experiments, in which the "A-robot" (named after Asimov) is charged with the mission to save human lives. The experiments include three different scenarios, where the baseline is set with A-robot only, and its task is merely to secure its own safety. In this scenario, robot A was able to avoid falling into a hole with 100% reliability. The second scenario added robot H representing human being, while the third scenario included two other robots, H and H2, representing two people. In the scenario when there was only one human life involved, the A-robot performed the task successfully. In the scenarios where there were two human beings (in the form of "H-robot") facing the danger of falling into a hole, in almost half of the trials, "the A-robot went into a helpless dither and let both 'humans' perish" (Deng 2016). Winfield et al believe that what they have set out in their experiments "match remarkably well with Asimov's first law of robotics" (Winfield et al 2014, 89). Thus, the A-robot's failure demonstrates the inadequacy of the Asimov's Laws in dealing with more complex ethical scenarios. To fix this problem, we would need to intro-

duce extra rules about how a robot can make choices in a moral dilemma.

Now we turn to the common proposals of ethical codes for robots: the Kantian ethics and Utilitarian ethics.

### *9.3.2.2 Kantian Ethical Codes for AMAs*

One of the leading ethical models for machine ethics is Kantian moral theory, which is regarded as “one of our best chances for the successful implementation of ethics into autonomous robots” (Tonkens 2009, 422). Kant’s normative ethics is a form of deontology, appealing to one’s sense of duty, not emotions, in making moral judgments. “Duties are rules or laws of some sort combined with some sort of felt constraint or incentive on our choices” (Johnson and Cureton 2017). It is understandable why Kant’s moral philosophy would be considered a natural choice for machine ethics: for Kant, human’s self-interest, desire, natural inclinations, moral sentiments such a sense of honour or the feeling of sympathy and compassion, all have absolutely no moral worth. According to Kant, even in a case where one finds an inner pleasure in spreading happiness around them, and takes delight in the contentment of others as one’s own work, however right and amiable this case may be, it “has still no genuinely moral worth” (Kant 1993, 11). Genuine moral acts must be done solely from the sense of duty of a rational being. Without the noises of human sentiments and desires, robots would be ideal for the implementation of Kant’s purely rational moral schema.

Kant’s first categorical imperative is stated as follows:

*[D1] I ought never to act except in such a way that I could also will that my maxim should become a universal law. (4:402) [Or: Act only in accordance with that maxim through which you can at the same time will that it become a universal law. (4:421)]*

A maxim is a person’s subjective volitional principle, while a universal law is a moral law that binds all rational beings in nature. Kant’s categorical imperative is based on the assumption that humans’ moral decisions come after deliberation of particular situations in accord with

the individual principle that one adopts for dealing with the situation at hand. As stated this way, Kant's categorical imperative demands both intrapersonal consistency and interpersonal agreement. This categorical imperative serves more as an injunction against acts that do not follow universalizable maxims, such as committing suicide because of misfortune, borrowing money with the knowledge that one can't pay it back as promised, not cultivating one's natural talents while indulging in the pursuit of pleasure, or not offering assistance to someone in dire need of help, when doing so would not put one in undue distress.

In terms of robotic ethics, we can have the moral rule formulated as such:

*[DR1] A robot should act only in such a way that the option chosen could in principle be a universal law for other robots.*

Since robotic ethics needs to be designed as situation ethics; that is, case-by-case deliberation on the course of action to take, [DR] focuses on the option rather than the maxim. The robot would have to be equipped with the function of gathering data on probable consequences and calculating the results in each case. In other words, the robot would have to be a consequentialist.

Whereas the first categorical imperative serves as an injunction, Kant's second categorical imperative gives a more specific moral guidance:

*[D2] Act in such a way that you always treat humanity, whether in your own person or that of another, always at the same time as an end, never merely as a means. (4:439)*

People must never be treated merely as a means to an end. People have their free will. To treat them merely as a means to an end is to deny their autonomy. In terms of robotic ethics, the categorical imperative can be formulated as such:

*[DR2] A robot must act in such a way that it always treats humanity never simply as a means, but always at the same time as an end.*

When humans are faced with the trolley problem dilemma, the typical response is that they would choose saving five lives and sacrificing one person in the standard scenario, but most would refuse pushing one person over the footbridge in order to save the five people on the track. In the trolley scenario, the rationale seems to be in favour of the 5:1 human life ratio. In the footbridge scenario, on the other hand, pushing someone over the bridge in order to stop the trolley is a clear violation of the second categorical imperative. According to Joshua Greene, “People exhibit a characteristically consequentialist response to the trolley case and a characteristically deontological response to the footbridge case.” (Greene 2007, 42) The reason for people’s different reactions, according to Greene, is that “the thought of pushing someone to his death in an ‘up close and personal’ manner (as in the footbridge dilemma) is more emotionally salient than the thought of bringing about similar consequences in a more impersonal way (e.g., by hitting a switch, as in the trolley dilemma)” (Greene 2007, 43). Therefore, Kant’s deontological approach is actually “a kind of moral confabulation” since deontological judgments are prompted by emotional responses (Greene 2007, 63). Greene calls this “the secret joke of Kant’s soul.”

According to Greene, when we make conscious moral decisions, we “respond to the conscious deliverances of our unconscious perceptual, mnemonic, and emotional processes by fashioning them into a rationally sensible narrative, and without any awareness that we are doing so.” (Greene 2007, 62) However, such an influence from the unconscious would not exist for robots and other forms of artificial intelligence. We can now consider whether Kantian ethics would compel robots to make the same decision that humans do.

In the footbridge scenario, the robot would obviously refuse to take the option of pushing the fat man over the bridge, since doing so would be a clear violation of [DR2]. So this is a no brainer. In the standard trolley scenario, however, the robot’s moral guidance is not so definitive. If the robot were to act on the principle that it judges to be universaliza-

ble for all robots, then it would likely be too paralyzed to do anything. Human beings will often have a spontaneous, intuitive judgement on whether their own maxims are suitable to be universal laws. Asking a robot to make such judgments, on the other hand, requires that the robot be equipped with either a huge databank containing all possible consequences for other robots to act in the same way, or the kind of intuition that humans have, but such intuitions are not available to artificial intelligence.

Herein lies the fundamental paradox of designing Kantian AMAs. Kant's first categorical imperative is based on his metaphysics of morals, according to which all human beings are autonomous rational beings. Rational beings are citizens of the Kingdom of Ends, sharing the same common laws and abiding by the same moral principles that they themselves legislate. Their autonomy lies in the sense that they are fully rational agents who "have an equal share in legislating these principles for their community" (Johnson & Cureton 2017). Artificial moral agents are, by design, machines that obey the programmers' orders. They do not act to legislate their own laws; they do not respect one another as equal law-makers. Furthermore, they lack freedom of the will, which is essential to the status of Kantian moral agency. For this reason, Ryan Tonkens calls them "anti-Kantian." Tonkens says, "Because we require our Kantian AMAs to act ethically, the fact that their development is a violation of Kantian morality renders their creation morally suspect, and our role as their creators somewhat hypocritical" (Tonkens 2009, 429). Therefore, not only is it problematic to implement Kant's ethical codes into artificial intelligence, we are actually acting immorally by creating Kantian AMAs. Tonkens argues, "By creating Kantian moral machines, we are treating them merely as means, and not also as ends in themselves. According to Kant, moral agents are ends in themselves, and because of this they ought to be respected as such. To violate this law is to treat an agent merely as an object, as something used for achieving other ends" (Tonkens 2009, 432-3). In other words, even if we could

create robots that would be able to follow our DR2, the very creation itself has already violated Kant's ethical principle.

### 9.3.2.3 *Utilitarian Ethical Codes for AMAs*

Applying utilitarian ethical codes for AMAs is another popular proposal. The utilitarian principle, simply put, is to judge the merit of an act by the potential consequences: actions are right insofar as they promote happiness or pleasure, and actions are wrong insofar as they generate pain. To John Stuart Mill, pleasure and pleasure alone has intrinsic value. Their value consists in people's desire. In other words, "good" is identified with "desirable." Mill says, "the sole evidence it is possible to produce that anything is desirable is that people do actually desire it" (Mill 2001, 81). What people desire would be good consequences; what people detest would be bad consequences. The most important feature of utilitarianism is the sole consideration of the number of people affected by the act instead of the self-interest of the actor. The utilitarian principle has standardly been formulated as follows:<sup>225</sup>

*[U] An act is right if and only if it produces a greater balance of good over bad in its consequences for all people affected, than any other act available to the agent.*

In the context of artificial intelligence, we can reformulate [U] as Utilitarian Robot [UR]

*[UR] In weighing the consequences of available courses of action, a robot must choose the option that will either generate the maximum benefits, or prevent the greater harm, for all human beings involved.*

According to Julia Driver, "Since the early 20th Century utilitarianism has undergone a variety of refinements. After the middle of the 20th Century it has become more common to identify as a 'Consequentialist'

---

<sup>225</sup> Utilitarianism can be act utilitarianism as formulated here, or rule utilitarianism: An act is right if it accords with a rule the general following of which produces a greater balance of good over bad for all people affected, than any alternative rule. Since artificial intelligence needs more precise rules and programs for each act, here we are only discussing act utilitarianism.

since very few philosophers agree entirely with the view proposed by the Classical Utilitarians, particularly with respect to the hedonistic value theory” (Driver 2014). We shall now look at some studies on the consequentialist model of machine ethics.

In a joint study conducted by Jean-François Bonnefon of the University of Toulouse, Azim Shariff of the University of Oregon, and Iyad Rahwan of MIT, participants were asked to evaluate autonomous vehicles that apply utilitarian ethical codes to favour sacrificing themselves and their driver to avoid running over a group of pedestrians. The study found that although participants approve of such vehicles for the greater good, they themselves would not want to buy this kind of automatic vehicle (AV). The study found that participants “overwhelmingly expressed a moral preference for utilitarian AVs programmed to minimise the number of casualties.”<sup>226</sup> However, when asked whether they would personally purchase such a utilitarian AV, participants were less positive. The researcher noted that “even though participants still agreed that utilitarian AVs were the most moral, they preferred the self-protective model for themselves” (Bonnefon et al 2016, 1574). This double standard creates a social dilemma: “Although people tend to agree that everyone would be better off if AVs were utilitarian (in the sense of minimising the number of casualties on the road), these same people have a personal incentive to ride in AVs that will protect them at all costs. Accordingly, if both self-protective and utilitarian AVs were allowed on the market, few people would be willing to ride in utilitarian AVs, even though they would prefer others to do so” (Bonnefon et al 2016, 1575). Without government regulation, such a utilitarian model of automatic vehicles would not be in the market; however, the possibility of government regulation would create even more resistance toward adoption of such a model. In other words, “regulating for utilitarian algorithms may

---

<sup>226</sup> “Overall, participants strongly agreed that it would be more moral for AVs to sacrifice their own passengers when this sacrifice would save a greater number of lives overall.” (Bonnefon et al 2016, 1574)

paradoxically increase casualties by postponing the adoption of a safer technology” (Bennefon et al 2016, 1573). The potential problem of a utilitarian model for artificial agents is manifested in this kind of conflicts between social utility and personal self-interest. If participants in these studies would be reluctant to purchase an automatic vehicle designed with the utilitarian model, then general public would be most likely resistant toward the idea of ethical robots implemented with the utilitarian ethics.

Other than the undesirability of utilitarian artificial intelligence, there is also the grave danger of having such artificial moral agents around in our society. The different responses people have with the trolley case and the footbridge case show that humans would refrain from certain actions that involve clear and personal harm. Humans do not always favour a utilitarian moral consideration, especially when the smaller number includes themselves, their kin and their acquaintances. Except for exceptional heroic acts, few people would willingly sacrifice themselves or their loved ones to produce the greater utility for all or for the greater good. Artificial agents, on the other hand, have no such inhibitions. Under [UR] and without any other overriding moral principle, they could undertake major destruction if doing so could lead to maximum benefits. This model may face the same difficulties that confront Asimov’s Zeroth Law.

The contrast between humans and AIs also demonstrates the fact that utilitarianism is never the go-to principle for humans’ moral deliberation. Even when we appeal to the principle of utility, our sentiments, self-interest, and other considerations would always make our utilitarian thinking “impure.” And yet if we do make purely utilitarian moral deliberation, as artificial moral agents would, then the outcome would be highly dangerous for human society. As Anderson & Anderson (2007) points out, utilitarianism “can violate human beings’ rights, sacrificing one person for the greater net good. It can also conflict with our notion of justice—what people deserve—because the rightness and wrongness

of actions is determined entirely by the future consequences of actions, whereas what people deserve is a result of past behaviour” (Anderson & Anderson 2007, 18). Utilitarianism may have its appeal in normative ethics only because humans do not abide by it completely and absolutely.

## 9.4 Confucian Robotic Ethics

Confucian ethics is not a form of rule-governed normative ethics. It is rather a form of virtue ethics, focusing on building the moral agent’s moral character and virtuous traits. To convert Confucian ethics into a form of implementable moral rule, we must do a liberal interpretation of the kind of Confucian moral rules that can be extracted from *The Analects*.<sup>227</sup>

There are many highly emphasised virtues in *The Analects* that can be formulated into moral rules for Confucian robotic ethics. I shall choose three main virtues: loyalty (*zhong*), reciprocity (*shu*)<sup>228</sup>, and humanity (*ren*), along with other supplementary virtues such as trustworthiness and righteousness. The first two are chosen in light of the comment of one of his chief disciples Zengzi (Master Zeng): what Confucius meant by “a single thread” of his Way is nothing but loyalty and reciprocity (*The Analects* 4.15). The third virtue is chosen because it is the overarching virtue in the whole Confucian tradition, underscored by both Confucius and Mencius and further elaborated in neo-Confucianism.

With regard to loyalty, Confucius has many things to say. It is one of Confucius’ four teachings (*The Analects* 7.25) and is said to be in conjunction with *shu* to form the single penetrating thread in his teachings (*The Analects* 4.15). In personal moral cultivation, Confucius says

---

<sup>227</sup> When not specified, the translations of texts in *The Analects* are mine with consultation of Dawson 1993 and Ni 2017.

<sup>228</sup> This is Peimin Ni’s translation. The word ‘*shu*’ is also frequently translated as ‘empathy’ (I have translated it as empathy in the past).

that a superior person (junzi) takes loyalty and trustworthiness (xin) as his first principles (The Analects 1.9; 9.25). In answering a student's question about how to promote virtue and discern delusion, Confucius' advice was: hold fast to loyalty and trustworthiness, and "move toward what is right—this is the way to promote virtue" (The Analects 12.10, Ni 2017, 288). In government, this virtue also plays an important role. Confucius told the ruler in the state of Lu that to gain loyalty from the people, the ruler himself must be "filial and caring" (The Analects 2.20). He told Duke Ding that the ruler must employ ministers with ritual propriety, while the ministers must serve the ruler with loyalty (The Analects 3.19); however, such loyalty is not blind obedience, but "offering counsel" to the ruler (The Analects 14.7). When a student asked about Prime Minister Ziwen, who "thrice took the office of prime minister and showed no joy in his countenance." He was thrice deposed from his position as the prime minister, but he showed no resentment. In addition, he always reported the previous policies to the succeeding prime minister. Confucius' evaluation of him is that he indeed has the virtue of loyalty (The Analects 5.19, Ni 2017, 160). When a student asked about governing, Confucius counselled: abiding in its affairs without weariness; conducting its affairs with loyalty (The Analects 12.14). The above quotes show that in Confucius' mind, loyalty is both a private and a public virtue: it is crucial both in one's conducting oneself and in one's engagement with public affairs.

In my analysis, "Loyalty is not a relationship directed toward others; rather, it is directed towards the role one plays. In this sense loyalty can be defined as 'doing what one is supposed to do' or 'being loyal to one's role.' In other words, a social role is not simply a social assignment; it is also a moral assignment. Being loyal to one's role means being able to act in accordance with whatever moral obligation that comes with the social role. Loyalty is thus being loyal to one's moral obligation and fulfilling the duty that one's role dictates" (Liu 2006, 50). This interpretation is further supported by Confucius' advice to a student, "loyalty in

relationships with others” (The Analects 13.19). The “others” here is not specifically one’s superior, but also one’s friends or strangers.

With this interpretation of the virtue of loyalty, we can now have our first moral principle for Confucian robotic ethics:

*[CR1] A robot must first and foremost fulfil its assigned role.*

Loyalty to one’s task is chosen to be the first law in Confucian robotic ethics because artificial intelligence, with its potential superpower, should particularly be designed to be role-specific and not omnipotent. Confucius says, “To guard Dao is not as good as to guard one’s station/role” (Zuozhuan Zhaogong 20). According to Kam-por Yu, “The Dao is of course the higher or final goal, but the question is whether everyone should aim at the Dao directly, or one should just fulfil one’s role faithfully, as the realisation of the Dao relies not just on one person, but on the collaboration of a number of people fulfilling their roles.”<sup>229</sup> Confucius also taught: “If one is not on the post, then one does not meddle with the managerial affairs” (The Analects 8:14). This comment demonstrates his view that one should do what is in one’s duty and not overtake someone else’s.

On first impression, this ethical code for artificial intelligence may seem trivial: of course machines are designed to complete the tasks. However, since we are discussing the possibility of future artificial moral agents that could make judgement calls under certain dire situations not previously anticipated by their programmers, we need to be prepared for these scenarios. [CR1] has a clear division of labour: a robot designed to offer health care should be specifically loyal to such a role, not to make other judgments such as whether the patient’s life is not worth keeping or to render assistance to meet the patient’s desire for euthanasia. An intelligent automatic vehicle should fulfil its duty to ensure safe driving for its driver, and thus ought not to take any act to sacrifice its driver by hitting against a tree in order to prevent a catastrophic disaster

---

<sup>229</sup> The quote is from personal communication.

for a school bus or the deaths of multiple pedestrians. It would be a mistake for us to try to design a “universal robot” as depicted in the play *R.U.R.*, where the word ‘robot’ was first coined.

An important virtue often paired with loyalty is reciprocity. Concerning reciprocity, Confucius commented that this word could “serve as guidance for practice during one’s entire life,” and he further defined it as such: “Do not impose on others what you would not wish for yourself” (The Analects 15.24, Ni 2017, 364). In a different passage, his chief disciple Zigong says, “If I do not want others to inflict something on me, I also want to avoid inflicting it on others” (The Analects 5.12, Dawson 1993, 17). From these two quotes, we can see that the connotation of *shu* is specifically defined as an interpersonal demeanour with a psychological preparedness. In contrast to the Christian Golden Rule: “Do unto others as you would have them do unto you”, this has often been called the negative Golden Rule, as it states injunction on what not to do, rather than specific commandment on what to do. In my analysis, this formulation of reciprocity is better than the Golden Rule in that what people do not desire, seem to have more common ground than what people do desire. In general, we do not wish for others to humiliate us, to deprive free will from us, to steal from us, to harm us, or simply to mistreat us in any way. “It is reasonable that we do not mistreat others in these ways either. And even if we desire others to act in a certain way towards us, the Confucian Golden Rule does not counsel us to act this way towards others too. It thus avoids the problem of subjective imposition of preferences that we see in the positive formulation” (Liu 2006, 55).

In terms of robotic ethics, however, we encounter the problem of the lack of desire in robots. If robots do not have any desires on their own, then how do they assess whether the consequences of their act would be what others (other human beings) would not want to be imposed on them? I think we can solve this problem by adding an algorithm of the scale of human preferences into the design. This scale of preferences can

be programmed as a machine “preference function” in the manner that Hilary Putnam suggested for functionalism: there should be a preference partial ordering and an inductive logic (i.e. the Machine must be able to “learn from experience”), some “pain sensors,” i.e., sensory organs which normally signal damage to the machine's body, or dangerous temperatures, pressures, etc., and that “the inputs in the distinguished subset have a high disvalue on the Machine's preference function or ordering” (Putnam 1967, 435). In this way, the artificial moral agent can assign a negative value to harm done to other human beings as well as a disvalue to its own damage. The second rule for Confucian robotic ethics could be stated as follows:

*[CR2] A robot should not act in ways that would afflict the highest displeasure or the lowest preference onto other human beings, when other options are available.*

As so formulated, [CR2] is still a negative injunction on what not to do. A general application of [CR2] is that a robot should never harm a person by choice, never inflict pain on human beings without due cause, never deprive someone of prized possession unless there are overriding considerations, and so on and so forth. This moral rule is similar to Asimov's First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm. However, this moral rule is much more flexible than Asimov's First Law in that there might be prevailing displeasure or negative preference that would outweigh the negative preference of harm. For example, radical injustice may be ranked as more undesirable than the possibility of physical harm. It is therefore conceivable that robots could be involved in insurgence against injustice and abuse if they have been designed with the appropriate role assignment and the right set of preference ordering.

The virtue of trustworthiness can be seen as an indispensable safeguard in robotic ethics, in that our robots are designed to serve various functions in our society and to interact with human beings. We need to be able to entrust our robots to perform the tasks assigned, not to deviate

from our expectations, and not to deceive us. Designing robots with this virtue in its default state is not to equip it with a Kantian prohibition: “Do not lie,”<sup>230</sup> since sometimes lying might be the more virtuous thing to do in a given situation. A deontological rule against lying can be formulated in contrast with this rule of trustworthiness:

[DR3] A robot must never lie. A robot must always give truthful answers.

In contrast, the implementation of the virtue of trustworthiness would be as follows:

[CR3] *A robot must always speak and act in a trustworthy way.*

The difference between having an absolute mandate: “do not lie,” and having the virtue of trustworthiness in the robotic default design, is that the former is written as an inviolable rule, while the latter is modelled after what virtuous people would do to be trustworthy. Of course, establishing a databank incorporating existing trustworthy behaviour in multifarious contexts would be a daunting task. However, with the hybrid approach, the robot would also be equipped with the ability to learn from previous samples and deduce its own trustworthy conduct in a new situation.

Next, the virtue of righteousness (*yi*) is crucial for moral agency, in that righteousness is almost interchangeable with morality: doing the right thing in the right context. Righteousness is one of the four cardinal virtues (humaneness, righteousness, propriety, and wisdom) that Mencius highly emphasised, and Mencius’ theory of the four moral sprouts (*siduan*) treats the fully developed moral sprouts as the four cardinal virtues. In the *Analects*, Confucius contrasts a virtuous moral agent, a superior person (*junzi*), with a petty person (*xiaoren*), in this way: “The mind of a superior person is preoccupied with righteousness, while the mind of a petty person is preoccupied with profit” (the *Analects*, 4.16). He also uses righteousness as the criterion for his own conduct and life’s

---

<sup>230</sup> Kant claims that lying is always morally wrong, no matter what the motive or the consequences might be.

choices (The Analects, 7.16). Confucius' disciple Master You says that one's "virtue of trustworthiness must accord with righteousness for one's words to be carried out" (The Analects 1.13). This shows that righteousness is a core virtue in Confucian ethics. In Chinese, the word for righteousness, *yi*, is homonymous with the word for appropriateness *yi*. By association, being righteous means doing things that are appropriate for the situation, and thus must be accompanied by cognition and judgment. Righteousness is thus an intellectual virtue. Confucian situation ethics is exemplified in the conception of this virtue: doing the right thing in the right situation. There is no universal principle that applies in all situations. Equipped with this virtue, our robot will be asked to make judicious decisions in the given situation, rather than following a universal mandate to act in the same way no matter what the situation is. It will be a flexible ethical design that allows the robot to assess the situation to choose the most appropriate action. Confucius says of a superior person: "The superior person (*junzi*) does not set his mind either for or against anything absolutely. He simply chooses to do what is right" (The Analects, 4.10); accordingly, our robot should be designed with the following principle:

*[CR4] A robot must be flexible in assessing the situation and make the decision most appropriate to the situation.*

Next, we come to the set of virtues serving as the basis for societal norms and personal sense of propriety (*li*): humility and respectfulness. In Chinese, humility (*gong*) and respectfulness (*jing*) are typically used together as a compound term, *gongjing*, and in the *Mengzi*, Mencius lists the heart of humility and respectfulness as one of humans' four "moral sprouts." In Mencius' view, humans naturally have this proclivity for societal cooperation and conformity to social norms. Cultivating this natural tendency into the virtue of propriety can serve as the foundation for society's codes of propriety such as rites, rituals, and etiquette. In the *Analects*, the two words of humility and respectfulness are not combined, but the two virtues are often mentioned together. Confucius

praises one student as “having humility in the way he conducts himself, and having respectfulness in the way he serves his superior” (The Analects, 5.25). He also explains to a student that the essence of the virtue of humanity includes “handling oneself with humility, handling affairs with respectfulness, and having loyalty when doing things for others” (The Analects, 13.20). Furthermore, Confucius says that a superior person must think about “being humble in one’s demeanour and being respectful in handling affairs” (The Analects, 16.10). Sometimes the virtue of respectfulness is also manifested in one’s attitude towards others (The Analects, 2.7; 4.18; 5.16; 5.17; 11.15). The above quotations show that the virtue of humility is mostly associated with one’s general demeanour, while the virtue of respectfulness is mostly associated with one’s attitude in dealing with particular people and affairs. These two virtues can be implemented in the default design of the robotic speech and demeanour. We can formulate the next two rules as follows:

*[CR5] A robot must always be humble in its demeanour and speech.*

*[CR6] A robot must respect the task at hand and be respectful to its interlocutor.*

With these two design guidelines, we will not have robots who would swear or curse, or robots whose behaviour would be flippant, insolent, dismissive, or defiant. Even though we cannot stop other humans from lacking these two virtues in their speech and act, we at least would have our artificial members always behave in accordance with propriety.

The last, but not least, virtue selected for Confucian robotic ethics is that of humanity (ren). According to Peimin Ni, “The term ren is central to Confucius’ philosophy. It appears 109 times in The Analects, and of the 499 selections in the book, 58 are devoted to this subject” (Ni 2017, 32). In Confucius’ assessment, the virtue of humanity is the hardest to cultivate. Even his prize student Yan Hui could only maintain this virtue in his heart up to three months, while the rest of the students could at most keep it for a day or a month (The Analects 6.7). When asked to

judge whether someone possesses this quality, Confucius rarely granted it even though he would acknowledge that the person has some other redeeming qualities (The Analects 5.5; 5.8; 5.19). At the same time, whether one can obtain this virtue is purely a matter of volition according to Confucius. He says, “Is humanity ever far away? If I want to achieve it, then I am already there” (The Analects 7.30). Confucius gives humanity the highest praise: He thinks that only the humane people are capable of loving and loathing people (The Analects 4.3), and if anyone is devoted to this virtue, she would never be doing anything bad (The Analects 4.4). If we say that Kant’s ideal is the Kingdom of Ends, then Confucius’ ideal would be the Kingdom of Humanity. Confucius talks about residing in the circle of humanity, surrounded by people of like virtue (The Analects 4.1). He also commented that humanness is a robust and resilient virtue that the superior people should never forsake even for the short duration of a mealtime. The superior people should never deviate from humanness even in moments of haste or times of duress (The Analects 4.5). Humanity is indeed the core virtue of Confucius’ moral teaching.

However, there are only a few passages where Confucius gave a definite description of what humanity is. When one student asked about ren, Confucius replied: it is to love people (The Analects 12.22). Confucius also commented that if those in the higher positions sincerely care for their kin, then the people will aspire to be humane (The Analects 8.2). When Yan Hui asked about ren, Confucius informed him that it is simply to restrain oneself in such a way that one conducts oneself completely in agreement with the rule of propriety (li): “Do not look when it is against propriety; do not listen when it is against propriety; do not say things when it is against propriety; do not act when it is against propriety” (The Analects 12.1). When another student asked about ren, Confucius gave the three requirements: “respectfulness in private life, reverence in handling business, and loyalty in relationships with others” (The Analects 13.19). Another time Confucius listed five virtues to explicate

ren: “respectfulness, leniency, trustworthiness, diligence, and beneficence” (The Analects 17.6). The most definitive explication of the virtue of ren comes from this passage in The Analects: “A person of humanity is someone who, wishing himself to be established, sees that others are established, and wishing himself to be successful, see that other are successful” (The Analects 6.30; Dawson 1993, 23). In other words, what the virtue demands is that the moral agent aid fellow human beings and other creatures in their quest for self-completion. There is a further restriction on what one aims to accomplish, however. Confucius says, “The superior person helps others to realise what is good in them, and he does not help others to bring to completion the bad qualities in them” (The Analects 12.16). That is to say, the Confucian ideal of ren is to aid others in becoming better people themselves, or we can say, to aid others to attain the state of ren. This virtue would be the most essential feature that we want to build into our robots.

Converting this virtue into a moral precept for artificial moral agents, we now have [CR7]:

*[CR7] A robot must render assistance to other human beings in their pursuit of moral improvement, unless doing so would violate [CR1] and [CR2]. A robot must also refuse assistance to other human beings when their projects would bring out their evil qualities or produce immorality.*

To render assistance means that the robot’s action is pursuant to a human being’s explicit request or command. In other words, the robot does not act on its own to decide what is good for the human subject or what the human subject ought to bring to completion. At the same time, being programmed with this moral rule, the robot would refuse to assist when the human command is for some evil doings. In this way, we not only have artificial moral agents that would not do things to harm human beings, we also have the safeguard against other humans’ using robots to accomplish their evil aims.

When discussing Confucian ethics, one cannot leave out the virtue of filial piety, since family ethics is core to Confucian ethics. However, I

argue that filial piety is not applicable to machine ethics. Filial piety is a role-specific virtue generated in a naturalistic human family structure. The relationships between humans and robots, even in an artificial family setup, is not bound by humans' family ethics. In particular, filial piety is a narrowly construed one-way devotion from children to their parents, and such devotion encompasses such virtues as loyalty, humility, respectfulness, and propriety. If we design robots with these virtues mentioned in [CR1] to [CR7], then they would be able to respond to human interactions with the right set of actions. There should not be a particular set of responses geared only toward the robots' designers or their adopted parents. Having this exclusive relationship between the designer/parent and the robot could conceivably create great danger to human society.

With the above list of virtues, we now have the rudimentary form of Confucian robotic ethics. How would a Confucian artificial agent act in the trolley and the footbridge scenarios then? In the footbridge scenario, a robot implemented with the Confucian ethical codes would never take the action to push the fat man off the bridge, because doing so would be a clear violation of Confucian ethical rules. In the trolley scenario, the judgement call is more complicated. If the robot is the driver of the trolley or a railroad worker, then its duty would dictate that it should pull the lever to cause the least harm possible among available options. If the robot is simply a passer-by, on the other hand, then according to [CR1], the robot is under no obligation to take any action, and under [CR2], the robot's preference would be inaction rather than action. Therefore, a passer-by robot should not take any action to divert the runaway trolley, even if doing so would reduce the number of casualties.

In a nutshell, given the trolley dilemma, a Confucian ethical robot would not pull the lever unless its particular role is the trolley driver or railway supervisor. Given the footbridge dilemma, a Confucian ethical robot would not push the fat man off the bridge to stop the trolley, no matter what its role is. The robot's decision would thus be different from

the intuitive choice of most humans, as it would not be affected by its unconscious emotional struggles that humans have in the footbridge case (see Greene et al 2001). A Confucian robot would not inflict harm or impose undesirable consequences on anyone by its action, even if its nonaction would not prevent such harm or undesirable consequences on others. In the foreseeable future when we do have self-regulating artificial moral agents in our society, we would want them to choose inaction over action, when both would lead to harm and undesirable consequences to human beings.

## 9.5 Conclusion

In their important piece on machine ethics, Anderson & Anderson (2007) write: “The ultimate goal of machine ethics... is to create a machine that itself follows an ideal ethical principle or set of principles; that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of actions it could take” (Anderson & Anderson 2007, 25). They also argue that “one of the advantages of working on machine ethics is that it might lead to breakthrough in ethical theory, since machines are well-suited for testing the results of consistently following a particular ethical theory” (Ibid.). In this paper, we have considered four such ethical models for machine ethics: Asimov’s Laws, Kantian Categorical Imperatives, the Utilitarian principle of utility, and the Confucian virtue ethics’ essential virtues. By comparing their solutions to the trolley problem and the footbridge dilemma, this paper argues that the Confucian model is superior to the other three ethical models. Of course, we do not design artificial moral agents merely to deal with the trolley problem and the likes. In many practical aspects, Confucian AMAs could be a welcome addition to human society. First of all, a Confucian ethical robot would be designed with specific job descriptions suitable for the role it is assigned—to render assistance for senior citizens, to provide health care for patients, to offer guidance for

customers, to navigate the car with safety, and so on and so forth. Its primary duty is role-bound; hence, any other decision it might make under special circumstances cannot violate its duty. Secondly, equipped with a carefully calculated preference ordering, a Confucian robot would not take any action that would cause the greatest displeasure (including harm) or the highly undesirable outcomes for other human beings. This principle is superior to Asimov's [A1], in that it both allows more dimensions of the consideration of negative values, and gives the robot more flexibility in weighing the permissible courses of action. It is also better than the Kantian or the Utilitarian principles in that this moral principle is based on the Confucian negative Golden Rule, and serves as an injunction against wrongful acts rather than a subjective volitional principle to take action. In the foreseeable future where we might have artificial intelligence taking things into their own hands, this principle can safeguard us from their making intentional sacrifices on any human being, no matter to what greater good they consider such actions would lead. With other interpersonal virtues such as humility, respectfulness, trustworthiness built in, a Confucian moral robot would behave in ways that help foster a civil society. With the intellectual virtue righteousness implemented in its decision processing, a Confucian moral robot would not be bound by inviolable commands to act, but would rather assess situations to calculate the most appropriate action to take in the given situation. Finally, a Confucian moral robot would be a humane robot: it would operate under the guideline to assist rather than obstruct in humans' endeavour to do good deeds, to become better people, and to build a better world.

One might question why we choose the Confucian model rather than other forms of virtue ethics. Other virtue ethicists would recommend such virtues as moderation, empathy, compassion, benevolence, friendliness, honesty, and so on and so forth. Those virtues are important for artificial beings as much as they are for human beings; however, the Confucian virtue ethics stands out from other forms of virtue ethics in

that it aims at cultivating “superior” beings above the masses. It is a form of moral elitist virtue ethics. We cannot manufacture superior human beings, but if we have much control in the design of artificial beings, then we would want them to be better than us at least in the moral dimension. We should not intentionally design robots to be just like us, with all our human foibles and moral failings. Robots with the most advanced artificial intelligence will far surpass human beings in their intellectual capacities. If they are not at the same time designed with superior moral attributes, then one day they might pose a great threat for humans’ wellbeing or even survival.

## 9.6 References

- Anderson, Michael & Susan Leigh Anderson (2007) “Machine Ethics: Creating an Ethical Intelligent Agent.” *AI Magazine* Vol. 28, No. 4: 15-25.
- Anderson, Michael & Susan Leigh Anderson (2006). “Machine Ethics.” *IEEE Intelligent Systems* 21 (4), 10-11.
- Bonnefon, Jean-François, Azim Shariff & Lyad Rahwan (2016). “The Social Dilemma of Autonomous Vehicles.” *Science* Vol. 352, Issue 6293, 24 June 2016, 1573-1576. DOI: 10.1126/science.aaf2654 Dawson, Raymond (Trans.)
- Confucius (551-479 BC). *The Analects*. New York: Oxford University Press, 1993.
- Deng, Beor (2015). “Machine Ethics: The Robot’s Dilemma.” *Nature* 523, 24–26 (02 July 2015) doi:10.1038/523024a.
- Driver, Julia (2014). “The History of Utilitarianism.” *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward

N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/win2014/entries/utilitarianism-history/>>.

Govindarajulu, Naveen Sundar, and Selmer Bringsjord, Rikhiya Ghosh, Vasanth Sarathy (2019). "Toward the Engineering of Virtuous Machines." AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society January 2019 Pages 29–35 <https://doi.org/10.1145/3306618.3314256>.

Greene, Joshua D. (2007). "The Secret Joke of Kant's Soul." In W. Sinnott-Armstrong (ed.), *Moral Psychology*, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development. Cambridge, MA: MIT Press. 35-79.

Green, Joshua D., and R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen (2001). "An fMRI investigation of emotional engagement in moral judgement." *Science* 2001 Sep 14; 293(5537):2105-8.

Greenemeier, Larry (2016). "Driverless Cars Will Face Moral Dilemmas." *Scientific America* June 23, 2016 (<https://www.scientificamerican.com/article/driverless-cars-will-face-moral-dilemmas/>)

Herkewitz, William (2016). "The Self-Driving Dilemma: Should Your Car Kill You To Save Others?" *Popular Mechanics* June 23, 2016. (<http://www.popularmechanics.com/cars/a21492/the-self-drivingdilemma/>)

Johnson, Robert and Cureton, Adam (2017). "Kant's Moral Philosophy." *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), forthcoming URL = <<https://plato.stanford.edu/archives/fall2017/entries/kant-moral/>>.

- Kant, Immanuel (1993). *Grounding for the Metaphysics of Morals* (1785). James W. Ellington (Trans.). Indianapolis: Hackett Publishing Company, Inc. 3<sup>rd</sup> Edition.
- Liu, JeeLoo (2006). *An Introduction to Chinese Philosophy: from Ancient Philosophy to Chinese Buddhism*. Malden, MA: Blackwell.
- Mill, John Stuart (2001). *Utilitarianism*. George Sher (Ed.) Indianapolis: Hackett Publishing Company, Inc. 2<sup>nd</sup> Edition.
- Ni, Peimin (2017). *Understanding The Analects of Confucius: A New Translation of Lunyu with Annotations*. Albany, NY: SUNY Press.
- Putnam, Hilary (1967). "The Nature of Mental States." Reprinted in Hilary Putnam, *Mind, Language, and Reality*, Cambridge: Cambridge University Press. 1975, 429-440.
- Pereira L.M., Saptawijaya A. (2007) "Modelling Morality with Prospective Logic." In Neves J., Santos M.F., Machado J.M. (Eds.) *Progress in Artificial Intelligence*. EPIA 2007. Lecture Notes in Computer Science, Vol. 4874. Springer, Berlin, Heidelberg.
- Tonkens, Ryan (2009). "A Challenge for Machine Ethics." *Minds & Machines* 19: 421-38.

Wallach, Wendell & Colin Allen (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Winfield, Alan F. T., Christian Blum & Wenguo Liu (2014). "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection." In *Advances in Autonomous Robotics Systems*, 15<sup>th</sup> Annual Conference, TAROS 2014, Birmingham, UK, September 1-3, 2014. *Proceedings*. 85–9.