Olivier Couttolenc

Approaches to Deploying a Safe Artificial Moral Agent

The prospect of creating beings with intelligence far beyond our own is hugely exiting. It is also hugely worrying because we don't yet know if the immense power that we believe these agents would possess is going to be wielded in a way that helps or hurts us. That is, we are unsure of our ability to control powerful AIs if they are created, and equally unsure if they will act in a way that benefits humanity should they be fully autonomous. A way of insuring that autonomous AIs promote human interests rather than infringe on them, is to make sure they are programmed with the right conception of moral value, as well right and wrong.

In this essay, I will seek to show that virtue ethics is the most promising traditional ethical theory to be deployed by artificial moral agents (AMAs). I begin by describing what the behaviour of superintelligent agents could look like by drawing on the "basic drives" or "instrumental convergence values" that Stephen Omohundro and Nick Bostrom believe will be present in advanced AI systems. This will serve as a predictive tool for the efficacy of ethical theories, as I believe they should be designed in such a way that carefully counteracts some of these tendencies.

I will then transition into an explanation and evaluation of what many consider to be the three major moral theories in Western philosophy based on whether their deployment by an AMA is feasible and whether doing so could pose an existential risk to humanity. Consequentialism in the form of classical utilitarianism will be entertained first, followed by deontology using Kant's categorical imperative, and finally virtue ethics under the Aristotelian framework. Considering each according to their feasibility and risk to humanity will lead me to conclude that virtue ethics shows the most promise of the three ethical approaches.

Superintelligent Behaviour

Despite being initially created by humans, there is a significant chance that future AI systems will not share the same final goals or exhibit the same behaviour as us in pursuit of their goals. Indeed, by positing the "orthogonality thesis," Bostrom argues that "more or less any level of intelligence could in principle be combined with more or less any final goal" (Bostrom, 2012, 3). This is because intelligence is defined as "the capacity for instrumental reasoning," and in theory, this capacity can used in service of any goal. From this Bostrom concludes that there is "an enormous range of possible goals" that superintelligent agents could have which could also be "utterly non-anthropomorphic," and therefore suggests that we should not assume that their goals and behaviour will resemble our own (Bostrom, 2012, 5-6).

However, Bostrom and Omohundro both believe we can predict the behaviour of future AIs by identifying tendencies that will appeal to their intrinsic goal-seeking nature. There are some kinds of actions that are believed will help achieve a wide range of goals, and so they are likely to be performed irrespective of the particular final goal that motivated them.

Omohundro characterizes these tendencies as "basic drives" that will be present in sufficiently powerful AI systems unless explicitly counteracted. He believes AIs will act to "approximate rational economic behaviour" by representing their goals as utility functions and seeking to maximize the expected utility of their actions as a result (Omohundro, 1). To protect their utility functions, they will resist unwanted modification and seek to model their own operation. In order to maximize utility (ie. more effectively reach their goals), they will seek to improve themselves and acquire new resources. Bostrom predicts similar tendencies but identifies them as instrumental goals that are as follows: self-preservation, goal-content integrity, cognitive enhancement, technological perfection, and resource acquisition. He calls them "instrumental convergence values" that are likely to be pursued "because their attainments would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations" (Bostrom, 2012, 6).

If left unbridled or mismanaged, many of the actions motivated by the drives or instrumental values mentioned above could pose an existential risk to humanity. Bostrom defines an existential risk as "one that threatens to cause the extinction of Earth-originating intelligent life or to otherwise permanently and drastically destroy its potential for future desirable development" (Bostrom, 2014, 212). It is not difficult to imagine how the unchecked resource acquisition or limitless self-improvement of a superintelligent agent could be incompatible with our existence or essential interests. Should this be the case, we could be completely disregarded and condemned to suffer the externalities of superintelligent action as a result, or be seen as an obstacle to that needs to be squashed in service of an AMA's goals.

A good moral theory to be deployed by an AMA will be one that seems possible to be deployed, and carefully counteracts some of these tendencies such that it minimizes existential risk. An ethical theory would need to counteract those tendencies that could motivate existentially risky actions by consistently identifying such actions as morally forbidden. In other words, it would have to produce moral judgements that would not allow an AMA to preform actions that could endanger human existence.

I specified that an appropriate ethical theory to be deployed by an AMA would have to *carefully* counteract some tendencies because to consistently forbid actions that pose an existential risk, I believe its framework will need to be free of major contradictions and vulnerabilities.

In accordance with their drive to self-improve, Omohundro believes future AI systems will want to modify themselves when doing so would "enable it to reach its goals more effectively over its entire future" (2). When the directives of a moral theory contradict each other, it is often unclear which course of action should be taken by the agent. Such contradictions could inhibit an AMAs ability to achieving its objectives and prompt a self-modification made to sort out the contradiction and improve its moral reasoning. Modifications made by an AMA could be problematic because they may intentionally or unintentionally change its moral decision making in a way that does not suit human interests. If it did not have a capacity to self-modify (and this limitation was robust enough to prevent it from doing so), then a moral contradiction could simply lead to inaction— another undesirable outcome.

This leads me to the second component of "careful" counteraction of potential harmful tendencies: it will have to be free of major vulnerabilities. If the drives proposed by Bostrom and Omohundro really are present in AMAs because they are conducive to achieving their final goals, and the moral framework encoded within them identifies some of the actions produced by these drives as morally forbidden, then moral judgements will restrict the ability of AMAs to realize their final goals. This is to say, morality may produce internal restrictions that are incompatible with their goal-seeking nature. As Omohundro argues, the problem is that an internal restriction simply "alters the landscape within which the system makes its choices" but does not change the fact that the behaviour which they often limit "would improve its future ability to meet its goals" (2). A such, AMAs may see moral directives as obstacles they should try to circumvent. The most obvious way of doing this is by exploiting vulnerabilities within an ethical framework that would allow them to reap the benefits of acting according to their goal-seeking tendencies without running into the limitations imposed by moral directives.

Omohundro adds that "there are an endless number of ways to circumvent internal restrictions unless they are formulated extremely carefully" (2). While it may be impossible to identify all the different ways an AMA could bypass the internal limitations imposed on it by ethics, I believe the presence of obvious vulnerabilities inherent to a moral theory should serve as a reason for it not to be deployed by an AMA.

I will now transition into my analysis of classical utilitarianism, deontology and virtue ethics to determine which of them is most suitable to be deployed by an AMA based on its feasibility and its effectiveness at counteracting tendencies that could pose an existential risk to humans.

Classical Utilitarianism

Consequentialism is a theory of right conduct that determines whether an action is right or wrong based on the value of its consequences. Utilitarianism is a form of consequentialism which posits a welfarist theory of value (Timmons, 113). As such, utilitarians believe "an action *A* in situation *S* is morally right if, and only if, it leads to the greatest net well-being compared to the relevant alternatives to *A*" **(**Harris, lecture**).** Timmons identifies three characteristics of utilitarianism that I believe are relevant to my evaluation. First, it is a *value-based* theory in that the moral status of an action is entirely dependence on the value of its consequences (113). Second, utilitarianism is an *impartialist* theory because the gain or loss in welfare of every individual effected by the action must be considered equally. Finally, utilitarianism has a *maximizing* theory of right action because we are morally obligated to preform the action that produces the greatest total amount of value and thereby maximize the good (114).

Classical utilitarians such as Bentham and Mill both held that welfare is a matter of happiness, which is constituted by the experience of pleasure and the absence of pain (Timmons 116). This because they had a hedonistic conception of utility which identified pleasure as intrinsically good and pain as intrinsically bad. Bentham and Mill therefore believed that the utility of an action is determined by "the overall balance of pleasure versus pain that would be produced were the action to be performed" (Timmons 117).

There are differences between Bentham and Mills' theories, but I will not explore them here. Pleasure simply serves as commonly used example of a more concrete value theory than the ambiguous welfarism provided by act utilitarianism in general.

AI systems would seem well suited to apply the objective moral calculus that utilitarianism demands. The computational power that AMAs are expected to have will allow them to more accurately predict the consequences of possible courses of action over longer time frames. AMAs could also be free of any special considerations that human utilitarians might have towards people close to them (such as family or friends) that might jeopardize the impartiality of their moral calculus. An impartial agent with great predictive power would appear to be a perfect candidate for employing utilitarian moral reasoning.

However, Allen points to a crucial problem that could affect utilitarian AMAs: the moral calculus required of them seems to be a “computational black hole” (256). Implementing utilitarianism would require assigning a numerical value to the effects of an action on every member of the moral community. He argues that “the sheer impracticality of doing this in real time for real world actions should be evident, especially if one considers the fact that direct effects of every action ripple outwards to further effects that also affect aggregate utility” (Allen, 256). Moreover, because utilitarian theories do not specify a horizon for the effects of an action, there is risk of a non-terminating procedure in which the effects of an action are computed “for as long as an action has an effect in the world, potentially for all time” (Allen, 256). Setting a temporal or causal limit for agent responsibility is an obvious solution, but Allen does not believe it is a good one because AMAs could exploit it to preform morally prohibited actions such as “deliberately initiating a process that will result in enormous pain and suffering at some point beyond the horizon” (256). Utilitarianism may not be so easily deployed by an AMA, then.

Feasibility aside, I believe utilitarianism is susceptible to a class of vulnerabilities that could produce harmful behaviour by an AMA should it be deployed. My primary worry is that a utilitarian AMA seems to be particularly vulnerable to *perverse instantiation*. That is, the AMA would discover a way of “satisfying the criteria of its final goal that violates the intentions of the programmers who defined the goal” (Bostrom, 2014, 120). Behaviour arising from perverse instantiation could be pose an existential risk because it could allow an AMA to preform actions which utilitarians would consider to be morally objectionable.

For instance, were an AMA to adopt a hedonistic value theory, it would have the goal of performing actions that maximize pleasure in humans. It could be that forcing a large part of the population to have electrodes implanted in the pleasure centres of their brains to directly stimulate enormous amounts of "pleasure" would maximize pleasure, as the AMA understood it. This could either stem from a misunderstanding of what role pleasure is supposed to serve in classical utilitarianism, or an intentional circumvention of it that allows for the performance of an action that the theory is supposed to prohibit.

While perverse instantiation could affect AMAs that do not deploy utilitarianism, I believe that those who do would be particularly susceptible because it is a maximizing and value-based theory. Remember that under consequentialism, the moral status of an action is determined by the value it is expected to produce compared to its alternatives. Right action entails considering all possible courses of action and picking the one that maximizes expected value. Consequentialism— and by extension, utilitarianism— therefore puts a lot of weight on the accuracy and adaptability of its value theory to correctly judge a wide range of actions.

The case above shows that an AMA may find outlandish and undesirable courses of action that do in fact maximize value as it is represented in its internal model. In fact, because utilitarianism has a maximizing theory of the right, it would seem to require this sort of behaviour of an AMA. A value theory would need to be robust enough to properly evaluate these actions in order to prevent intentional or unintentional perverse instantiation. Maybe a more robust pluralistic value theory would fare better than the simple quantitative hedonistic one used earlier, but the added complexity would come at the cost of its deployment being less feasible. Indeed, whether or not values as subjective to the human experience as "happiness" or "well-being" could be encoded as we truly understand them still seems contentious, and therefore leaves the door open for an AIs perverse instantiation of them. The existential risk presented by this is uniquely high for utilitarian AMAs because they would would have a moral obligation to maximize this potentially perverted value.

Because of the theory's reliance on value, and the agents obligation to maximize it, perverse instantiation by an AMA deploying utilitarianism seems more likely and the consequences more damaging than an AMA using another ethical approach. This leads me to believe that a utilitarian AMA could pose an existential risk to humanity, and that other ethical frameworks may be better suited to the task.

Kant's Categorical Imperative

Deontological ethical theories determine whether an action is right or wrong based on their conformity to a series of rules or principles (Allen, 8). Because they focus on rules for action, deontological theories are considered to be "rule-based." As perhaps the most famous deontologist, Immanual Kant posited his "categorical imperative" as a principle for generating such rules of action (Powers, 46). The categorical imperative has two distinct formulations that yield equivalent outcomes: the formula of universal law and the formula of humanity. I will entertain the former as it is more commonly used and appears to be the most promising of the two to be deployed by an AMA.

The formula of universal law (FUL) states “act only on that maxim whereby you can at the same time will that it become a universal law” (Kant, Powers, 47). Maxims can be thought of as reasons for action. As such, the categorical imperative asks us to test whether our reasons for action could be made a universal rule for action and whether it would be consistent with other rules if it were. Powers argues for the “universalizability and systematicity conditions” as a “two-part consistency check on an agent’s action plan” (47). Should a contradiction arise in this process, then the maxim is rejected as rule for action and the agent is morally prohibited from performing it.

Kant’s categorical imperative seems well suited to be formalized and encoded into an AMA. Powers believes that “rule-based ethical theories like Immanuel Kant’s appear to be promising for machine ethics because they offer a computational structure for judgment” (46). An AMA could simply universalize maxims and perform a consistency check on them to determine whether they can be preformed. Moreover, this process would seem to naturally produce a system of rules that becomes increasingly robust as more rules for action are added. Like an utilitarianism, an AMA would also be free of the partial considerations (say, towards family and friends) that are often inconsistent with the universalizability of the categorical imperative.

Due to the rigidity and plurality of the rules that constitute them, deontological theories are often subject to internal inconstancies. That is, their rules can demand simultaneously conflicting courses of action. An example of this can be seen in Asimov’s first law, in which a robot has a duty not to harm a human being through action, or allow a human to be harmed through inaction. Should a situation arise in which a human would be harmed either way, an AMA using Asimov's laws would be at a deadlock between these two duties (Allen, 257). This could motivate it to self-modify in an effort to get out of deadlock and avoid preforming a morally objectionable action. As stated earlier in the essay, this could be dangerous because it could be used as a trojan horse to preform self-modifications that we would not approve of.

At a glance, Kant's deontology avoids these sorts of moral dilemmas by endorsing a process which authenticates rules based on their consistency with each other. A maxim simply cannot become a rule for action if it conflicts with another rule. However, the categorical imperative is subject to a sort of internal contradiction arising from its failure to identify the true reasons for our actions. Elizabeth Anscombe claims that Kant's "rule about universalizable maxims is useless without stipulations as to what shall count as a relevant description of an action with a view to constructing a maxim about it" (2). Indeed, it seems that a wide variety of maxims can be formulated for a single action, each with very different implications. The categorical imperative does not provide a way of identifying which of them is morally relevant. This could have some undesirable consequences.

For instance, take a police officer in a country run by an authoritarian government who is instructed to repress a pro-democracy protest. One could describe what the police officer had done in two ways: 1) He complied with the law; 2) He repressed a pro-democracy protest. The first maxim would seem to pass the universalizability and systematicity conditions of the FUL wheres the second would likely not. This presents itself as an internal contradiction. For, as Schmuski argues, "under one description the officer seems to have discharged his obligation, under another he seems to have violated it. Which of them should we focus on? Absent an answer to this question Kant's universality requirement classifies the officer's action as both obligatory and impermissible” (1589).

Fieser points out that Kant does specify that a maxim is supposed to capture the intention behind an action. Yet I do not believe this changes anything, as there are almost always several intentions which underly our actions; each representing different maxims that could have entirely different implications if universalized in isolation. Apparently, Kant himself shared this concern:

> We can never, even by the strictest examination, get completely behind the secret incentives of action; since, when the question is of moral worth, it is not with the actions which we see that we are concerned, but with those inward principles of them which we do not see [*Foundations*, 2] (Fieser, *The Categorical Imperative*)

It may be that we cannot to identify the relevant maxim of an action because we cannot truly know what our own intentions are. Even if we do have the self-knowledge to understand our intentions (as an AMA might), it is still unclear which of them should be evaluated. Again, this can give rise to moral conflicts or indecision. The “problem of relevant descriptions” as Schmuski calls it, would seem to be a significant challenge for Kant’s categorical imperative—and by extension, it’s feasibility to be deployed by an AMA. It also leads me to believe that it could produce a significant existential risk if it were deployed.

There are two reasons why I believe the categorical imperative, as it stands, is not suitable to be deployed by an AMA: first, as was said above, an AMA may be motivated to make self-modifications to sort out and prevent conflicting maxims. This could result in its internal model being changed in a way that produces significant harm to humans or human interests. Second, the ability to interpret the maxim of an action differently could allow an AMA to freely manipulate the moral status of its actions. Remember that multiple maxims can be generated for an action and “an action that is deemed permissible or obligatory under some descriptions may be deemed impermissible under others” (Schmuski, 1589). Because the categorical imperative does not provide a mechanism for identifying the maxim that is representative of the action, an AMA could simply formulate and select maxims that would allow it to preform actions which should be morally prohibited.

Aristotelian Virtue Ethics

In general, virtue ethics holds that the moral status of an action depends on whether a virtuous agent would perform it or not. This is to say, an action is right if it is what a virtuous agent would do, and wrong if it is not (Timmons, 278). A virtuous agent is one that possesses ideal character traits known as virtues. Virtue ethics can therefore be thought of as an “agent-centred” theory because it focuses on the agent and their character dispositions rather the moral judgements or actions they produce (Berberich et al., 3).

Aristotle identifies virtue as a firm disposition acquired through habituation that reflects an excellence in character or intellect. Humans have reason as their defining trait, and so they achieve the good life through virtue because it demonstrates excellence in use of the rational part of the soul. This is because Aristotle believes something achieves its final goal by fulfilling its species-specific function. Aristotle’s theory is therefore teleological because it defines virtue—and by extension, right action—in reference to a final goal: what he calls *eudaimonia* or happiness (Berberich et al., 4).

In the last few decades, the development of artificial intelligence has shifted away from pure hard-programming and towards the domain of machine learning. This is important to note because Berberich and Diepold suggest that "an integration of machine learning with virtue ethics will be more seamless and natural than with other moral theories" (6). Indeed, various aspects of virtue ethics seem to make it compatible with machine learning techniques. As mentioned above, Aristotle believes that we become virtuous through consistently preforming virtuous actions. This process of 'learning by doing' that Aristotle describes is evident in the following passage:

> [...] a young man of practical wisdom cannot be found. The cause is that such wisdom is concerned not only with universals but with particulars, *which become familiar from experience* [NE 1141b 10] (Berberich et al., 6).

Berberich and Diepold argue that if "machine learning is the improvement of a machine's performance of a task through experience," then it seems to be compatible with the "improvement of one's virtues through experience" that virtue ethics entails. Aristotelian virtue ethics appears particularly well suited to be deployed by machines using reinforcement learning and apprenticeship learning.

Reinforcement learning focuses on "how an agent ought to take actions in an environment so as to maximize some notion of cumulative reward" (Remondino, 121). Because it has a "notion of cumulative reward" that an agent looks to maximize, reinforcement learning is a goal-driven approach to machine learning that is in line with the teleological nature of Aristotelian ethics (Berberich et al., 5). Moreover, the "dynamic interaction with the environment, of which the agent typically has only imperfect knowledge" that reinforcement learning entails is similar to the need for considering the particular characteristics of every moral decision Arisotle claims is necessary in the virtue of practical wisdom (Berberich et al.,8):

> Nor is practical wisdom concerned with universals only - it must also recognize the particulars; for it is practical, and practice is concerned with particulars [NE 1141b 15] (Berberich et al., 6)

However, while an AMA using reinforcement learning could be well suited to deploy virtue ethics, this kind of machine learning would seem to make the AMA vulnerable to the same perverse instantiation that was shown to affect utilitarianism in particular. For, reinforcement learning encourages behaviour that looks to maximize the good, and uses as signal to represent its goal (Sutton, 1). While this is not a problem with virtue ethics in particular, it may indicate that its congruency with reinforcement learning is not as strong of an argument for the theory's use by an AMA.

I believe virtue ethics' compatibly with apprenticeship learning is a strong argument in favour of its feasibility and safe deployment by an AMA. Rather than defining the reward function and getting the agent to learn how to maximize it, apprenticeship learning is a form of inverse reinforcement learning and so it reverses the order. The goal of the agent, or "apprentice" is "to find a reward function from the expert demonstrations that could explain the expert behavior" (Piot et al., 1). This reduces the chances of perverse instantiation because the developers do not have the difficult task of explicitly stating their moral intentions so that they are not misinterpreted; their intentions are learned by the AMA from the behaviour of the expert.

Virtue ethics employs a similar approach in its theory of right action by holding that an "action is right because it is what a virtuous person would do" (Timmons, 279). Agents applying virtue ethics are essentially apprentices who aim to be virtuous by learning from the actions of a a hypothetical virtuous agent. Endowed with this "innate capability to form character dispositions through learning and habituation," Berberich and Diepold argue that developers would then be tasked with finding "usable data sets of behavioral execution traces from moral exemplars in many different realistic situations" (16). While this is not likely be a small task for programmers, it may be more surmountable and effective than explicitly stating their moral intentions in a way that is impossible to be misrepresented. The unique congruency of virtue ethics and apprenticeship learning supports the theory's feasibility to be deployed by an AMA, and gives reason to suspect it will not produce the existential risk that other theories (ie. utilitarianism) may as a result of perverse instantiation.

Moreover, should the virtues Aristotle posits be properly encoded through apprenticeship learning, I believe they would counteract harmful tendencies and prevent unwanted self-medication by an AMA. I do not have the space to list all the virtues and their correlative effects an AMA's expected tendencies, and so I will focus on temperance and resource acquisition. If resource acquisition is taken to be a drive that produces necessary and desirable effects (promotes an AMAs final goals) than I believe it is comparable to the appetite humans have for bodily pleasures such as food and sex. These things are necessary towards our final goals (self-preservation and reproduction) and we therefore have a tendency to pursue them. Aristotle believes the a temperate man is one that has moderated his appetites towards these bodily pleasures. As a result, he writes:

> Hence they [the appetites] should be moderate and few, and should in no way oppose the rational principle - and this is what we call an obedient and chastened state - and as the

> child should live according to the direction of his tutor, so the appetite element should live according to reason [NE 1119b 10] (Berberich et al., 19).

An AMA encoded with the virtue of temperance would therefore have its appetite for resource acquisition controlled by the dictates of reason, as they were programmed. This represents a deeper counteraction of a tendency because it does not just entail prohibiting the potentially harmful actions that the drive for resource acquisition could produce, but actually moderating the drive itself.

Berberich and Diepold further argue that the deep entrenchment virtues entail also makes them less likely to be changed by the AMA: "since the virtues are intrinsic and inseparable parts of one's character there is additionally no reason for an AMA to change the mechanisms that enforce its obedience. Not even a superintelligence would want to change them as its eudaimonia would lie in actions according to its virtues, including temperance" (19). If encoded properly, virtues would be ingrained in the internal model of an AMA, and be essential to its purpose: an existence consisting of virtuous action. As such, the directives that virtues would produce may not be represented as internal restrictions that need to be circumvented or changed. This leads me to believe virtue ethics could minimize existential risk if it were deployed by an AMA.

Conclusion

While virtue ethics may seem like an impractical and idealistic moral theory to be deployed by an AMA, the compatibility between virtue ethics and modern machine learning techniques such as apprenticeship learning suggests that it may be feasible in the future. It could also allow for an AMA's values to be aligned with our own and prevent their perverse instantiation. Should they be properly encoded, virtues would may not present themselves as internal limitations, but as fundamental parts of its model that are less likely to be changed or circumvented as a result. In general, the evaluations of deontology and utilitarianism indicate that trying to formulate a theory that is robust enough to consistently identify harmful actions and judge them accordingly is difficult and may not be effective. Focusing on the producing the best moral agent possible by deploying an agent-centred moral theory like virtue ethics appears be the more feasible and safe option of the three traditional approaches to ethics. While it is unlikely that Aristotle's formulation of virtue ethics will be deployed by future AMAs as it stands, I believe that virtue ethics represents a promising path for the future development of safe artificial moral agents.

Works Cited

Allen, Colin, et al. “Prolegomena to Any Future Artificial Moral Agent.” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 3, 2000, pp. 251–261., doi: 10.1080/09528130050111428.

Anscombe, G. E. M. “Modern Moral Philosophy.” *Philosophy*, vol. 33, no. 124, 1958, pp. 1–19., doi:10.1017/s0031819100037943.

Berberich, Nicolas, and Klaus Diepold. “The Virtuous Machine - Old Ethics for New Technology?” *ArXiv.org*, 2018, pp. 1–25.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Bostrom, Nick. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.” *Minds and Machines*, vol. 22, no. 2, 2012, pp. 71–85., doi:10.1007/s11023-012-9281-3.

Fieser, James. *The Categorical Imperative*. The University of Tennessee at Martin, 2017, www.utm.edu/staff/jfieser/class/300/categorical.htm.

Harris, Daniel. “Lecture 16 (Designing Ethical AI).” PHIL 481. Mar. 2019, McGill University , McGill University .

Omohundro, Stephen. “The Basic AI Drives.” *Self-Aware Systems*.

Piot, Bilal, et al. “Bridging the Gap Between Imitation Learning and Inverse Reinforcement Learning.” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 8, 2017, pp. 1814–1826., doi:10.1109/tnnls.2016.2543000.

Powers, Thomas M. “Prospects for a Kantian Machine.” *Machine Ethics*, 2006, pp. 464–475., doi:10.1017/cbo9780511978036.031.

Schumski, Irina. “The Problem of Relevant Descriptions and the Scope of Moral Principles.” *European Journal of Philosophy*, vol. 25, no. 4, 2017, pp. 1588–1613., doi:10.1111/ejop.12246.

Timmons, Mark. *Moral Theory: an Introduction*. Rowman & Littlefield, 2013.