The Perfect Moral Agent,

Kantian artificial moral agency re-examined.

Master thesis by Jip Maat,

Student number: 566634jm

15 EC

10-07-2021

Written under supervision of Dr. Paul Schuurman

Thesis advisor: Prof.Dr. Payal Arora

22.675 words.

ABSTRACT: This thesis is concerned with the application of Immanuel Kant’s (1724-1804) moral philosophy to the newly emerging field of machine ethics. Specifically, the question that I aim to answer here is to what extent it is possible for machines to engage in genuine moral behaviour from a Kantian point of view. In order to answer this question, I review and evaluate several works from machine ethicists that have approached the question of machine morality from a Kantian perspective. Having established the most common strategies and problems in this field, I propose three domains in which additional work is necessary for the achievement of a Kantian artificial moral agent (AMA): machine rationality, machine autonomy and the ethical consistency of Kantian AMA’s with Kantian morality in general. I offer strategies for solving the problems in these three areas based on a close reading of Kant’s practical philosophy and earlier work done in machine ethics, and conclude that we may be fairly optimistic about the prospects of a Kantian AMA, on the condition that such AMA’s are built with the aim of furthering morality.

Table of Contents

[1. Introduction to machine ethics and Kant’s practical philosophy 1](#_Toc76819797)

[1.1 On machine ethics and its relevance 1](#_Toc76819798)

[1.2 Three schools of machine ethics 4](#_Toc76819799)

[1.3 Kant’s practical philosophy and its rational nature 6](#_Toc76819800)

[2. Attempts to conceptualize a Kantian machine, successes and failures 10](#_Toc76819801)

[2.1 Powers’ three approaches to applying a formalized CI 10](#_Toc76819802)

[2.2 Ulgen’s Kantian Machines 14](#_Toc76819803)

[2.3 Other approaches and considerations from machine ethicists 17](#_Toc76819804)

[2.4 The challenges for a Kantian machine 20](#_Toc76819805)

[3. The categorical imperative and the reasonable will 23](#_Toc76819806)

[3.1 The CI: pure analysis vs simulation 23](#_Toc76819807)

[3.2 Having a will, perfect and imperfect rationality 28](#_Toc76819808)

[3.3 Implicit Kantian agency in machines 34](#_Toc76819809)

[4. Freedom and autonomy 36](#_Toc76819810)

[4.1 Kantian autonomy 36](#_Toc76819811)

[4.2 Kantian free will and its compatibility with determinism 42](#_Toc76819812)

[5. The ethics of machine ethics 52](#_Toc76819813)

[5.1 The challenge unfolded 52](#_Toc76819814)

[5.2 The challenge accepted 55](#_Toc76819815)

[6. Conclusion 62](#_Toc76819816)

[Bibliography 65](#_Toc76819817)

# Introduction to machine ethics and Kant’s practical philosophy

## 1.1 On machine ethics and its relevance

In 1967, moral philosopher Philippa Foot introduced the world to a type of thought experiment that has gone on to influence perhaps every introductory class in moral philosophy since.[[1]](#footnote-1) In these ‘trolley cases’, we are asked to imagine a runaway trolley that is on its way to kill five innocent people that are tied up on the railway tracks just a bit ahead. We have the power to change the trolley’s direction however, the experiment states, by pulling a lever that will make our runaway trolley switch to a railway track that only has one person tied to it. By pulling the lever then, we save the five people tied up to the tracks, but kill the sixth person, who we assume to be just as innocent as the other five. The point of this experiment, obviously, is to question our moral intuitions. Should we minimize the loss of human life by pulling the lever, or let the trolley run its course so as not to get blood on our own hands? Does it make a difference if instead of pulling a lever, we are asked to push an extremely fat person in front of the trolley to stop it, or if we personally know any of the people on the tracks? Similarly, what are we to do if in the middle of a car accident, we face the choice between turning left and hitting two old ladies, or turning right and hitting a child? With the advent of modern driverless trains and self-driving cars - see for example the London Underground and the Waymo one, a driverless taxi service in Phoenix Arizona - a renewed interest in trolley problems has emerged. Though such autonomous vehicles cause less accidents than human drivers do, accidents in which the vehicles’ decision making has distinctly moral consequences are bound to happen. The existence of autonomous vehicles raises a host of difficult moral questions beyond our original question of how to act in difficult moral situations. If an autonomous car ‘decides’ to sacrifice a child in order to save two old ladies, for instance, and the human bystanders disagree with this choice, then who are these people to blame? Does the moral responsibility for this choice lie with the engineer, the programmer, the CEO of the car company or perhaps the car itself? Additionally, if there are two companies producing autonomous cars, do we demand the cars of both these companies to always make the same choice in the same situations, or is some amount of moral diversity allowed? We know, from the fact that trolley cases are still seen as relevant by moral philosophy teachers, that there exists no consensus among humans on how to deal with such dilemmas, so why should we demand our cars to always agree on questions about morality? Now in addition to autonomous vehicles, our increased reliance on technology has given rise to a number of situations in which the behaviour of this technology could raise ethical questions. Consider, for example, the advent of autonomous weapons systems in various countries' militaries, the implementation of robot caregivers in nursing homes or the automation of credit card and banking systems. In all of these contexts, humans are gradually handing over various amounts of autonomy to machines, which are being given charge over decision making processes that can have severe consequences for human actors.

The examples above illustrate the motivation for and domain of machine ethics, a relatively new field of study concerned with the ethics of machine behaviour towards humans and other machines.[[2]](#footnote-2) This field, in contrast to neighbouring fields of study concerning the ethics of human use of computers or the ethical implications of new sorts of technology in general, explicitly engages with the morality of the behaviour of machines themselves. This ethical investigation of machine behaviour, though a common theme in 20th century science fiction,[[3]](#footnote-3) only really became a serious philosophical discipline in and of itself in the early 2000’s, when the Association for the Advancement of Artificial Intelligence (AAAI) started organizing conferences on the subject. In the past 20 years, the field of machine ethics has expanded, gained political and public interest and developed some distinct branches and traditions. Much like in traditional ethics, the question of how machines should become ethical exactly has given rise to three main branches in machine ethics: deontological, utilitarian and virtue based machine ethics.[[4]](#footnote-4) In addition to these three directions of morality itself, machine ethicists disagree on the amount of autonomy and the sort of agency that moral machines are to enjoy.

James H. Moor, in *The nature, importance and difficulty of machine ethics*, distinguishes three ways of conceptualizing in what ways machines can be said to be moral.[[5]](#footnote-5) Implicit moral agency, first of all, is achieved when the actions of a machine conform to certain moral standards as evaluated by human observers.[[6]](#footnote-6) Any machine that is built to avoid behaviour that its designers consider to be ethically undesirable then can be said to be an implicit moral actor: it is so constrained that its actions are implicitly ethical. Explicit moral agency, in contrast, demands not only that a machine acts in conformity with some ethical principle, but that it is able to represent this principle itself as well. In this way, a machine can be able to justify its specific actions by showing how it has arrived at the decision to act in this way from its moral principle. Full moral agency, finally, consists of the ability to be held accountable for one’s actions. If an implicit or explicit moral machine goes awry, the ultimate responsibility for this fault lies with its programmers, as they are the ones who have programmed the machine to act in this way. A full moral agent (which according to Moor, only human adults are) is responsible for its actions, as it makes explicit moral judgements by itself and is competent enough to reasonably justify those actions.[[7]](#footnote-7) There is no agreement among machine ethicists as to what sort of agency machine ethics should aim for. Where some authors argue that only implicit moral agency is desirable, others maintain that only a machine with full moral agency can be said to be moral at all. In addition to this, scholars disagree as to whether or not explicit or full moral agency is even technologically possible. It is important, therefore, to specify what sort of agency we are talking about when we are discussing specific machine ethicists or theories, and in the following chapter I shall aim to do so.

## 1.2 Three schools of machine ethics

Utilitarianism, the school of ethics often associated with Jeremy Bentham and John Stuart Mill considers the ethical value of action to be dependent on that action’s consequences. That action that results in the greatest amount of utility or happiness for the greatest number of people is considered the best action, and in our ethical deliberation we should therefore evaluate the most likely consequences of our potential actions. Utilitarian approaches to machine ethics, such as Cloos’ utilibot project,[[8]](#footnote-8) try to incorporate this ethical principle by having machines compute the most likely outcome of a set of possible actions, and evaluate those possible outcomes so as to maximize the utility of their actions.[[9]](#footnote-9) There is, however, a set of standard objections to utilitarian machine ethics that is yet to be overcome by its proponents. First, it is as of yet unclear if the process described above is computationally attainable in the foreseeable future, since there are no theoretical limits to the amount of consequences that a machine would have to be able to evaluate in order for its choice to be sufficiently ethical from a utilitarian standpoint.[[10]](#footnote-10) In addition to this objection, the cold rationality of machines making invasive decisions based on calculated future benefits seems to some people to represent the very danger that machine ethics was meant to avoid.[[11]](#footnote-11) This might not seem like a large problem in the autonomous vehicles discussed above, but as we become more and more dependent on increasingly autonomous technology, the potential dangers of utilitarian machines become larger as well. Related to this objection is the fact that utilitarianism in general has been accused of not sufficiently guaranteeing the rights and integrity of individual people when these rights conflict with projected utility outcomes.[[12]](#footnote-12) For these among other reasons, the utilitarian approach to machine ethics, though far from abandoned, is not too popular nowadays.

The virtue based approach to machine ethics, commonly understood to have its roots in the work of Aristotle, emphasizes the importance of the development of moral character over the morality of individual actions. Virtue based approaches to machine ethics then often conceive of ethical machines as bottom-up, learning algorithms that aim to continuously better themselves through their moral experiences.[[13]](#footnote-13) As this approach does not specify an a priori overarching moral principle however, critics have pointed out that the bottom-up development of morality in machines could very well have dire consequences: what is to guarantee that such machines will develop values and virtues that humans agree on as being morally right? If we leave the development of ethics in the hands of machines themselves, the argument goes, it is not unlikely that the result could conflict with morality as conceived of by humans.[[14]](#footnote-14) It is perhaps because of these reasons that the virtue approach to machine ethics to this day remains underdeveloped when compared to its utilitarian and deontological counterparts.[[15]](#footnote-15)

Deontological ethics, finally, revolves around the formulation of a priori rules for conduct - duties - that, unlike utilitarianism, do not directly depend on the outcome of that conduct. Perhaps the most well-known deontological principle of machine ethics is science-fiction author Isaac Asimov’s formulation of the three laws of robotics, which specify a set of hierarchical rules for machine behaviour ultimately resting on the principle that machines are never to harm humans in any way.[[16]](#footnote-16) As Asimov himself showed in his numerous novels however, a potential problem for such an approach is the risk of encountering situations in which the moral duties of a machine conflict with each other, and result in unpredictable behaviour. A more promising deontological approach to machine ethics, one that has been considered by a number of authors that will be discussed in the next chapter, is based on the deontological moral philosophy of Immanuel Kant (1724-1804). Kant’s ethical system, which I will further explain shortly, emphasizes the rational nature of moral behaviour. This rational aspect of Kantian moral philosophy has made it appealing to machine ethicists in particular, as machines can be considered the ultimate rational beings.[[17]](#footnote-17) As this thesis is concerned with the Kantian approach to machine ethics, I will leave the utilitarian and virtue based approaches for what they are here and seek to establish some of the basic tenets of Kant’s moral philosophy so that we can apply these to the field of machine ethics in the following chapters.

## 1.3 Kant’s practical philosophy and its rational nature

Kant’s practical or moral philosophy is based on a supreme moral principle - or law - that he calls the categorical imperative (CI). This principle, in Kant’s view, can be characterized as an objective, unconditional principle that is purely rational, in the sense that it is both derived from the analysis of practical reason and that it does not rely on empirical data to specify what actions are and are not morally good. Because it is devoid of any empirical a posteriori considerations then, the CI can be said to be good in and of itself, in the sense that its being good is not dependent on any external conditions. Since the CI is the only principle that is purely rational in this way, only it can be the basis of morality. I will come back to this relationship between morality and pure rationality later in this chapter, but first let me show what the CI is exactly, and how we ought to put it to use.

In the *Groundwork to the metaphysics of morals,[[18]](#footnote-18)* perhaps the most fundamental of Kant’s practical works, the CI is formulated in three distinct ways, that all express the same principle and classify the same actions as good or bad, but that highlight different aspects of that principle. The first of these formulations, which explicitly focuses on the form of the moral principle, goes as follows: Act only on that maxim through which you can at the same time will that it should become a universal law.[[19]](#footnote-19) Maxims, for Kant, are subjective determinations of the will, or ‘principles of intention’.[[20]](#footnote-20) For any action that I consider doing, I should be able to specify a maxim that expresses both what I aim to do exactly and the reason for which I aim to do this. Maxims take the following form.  ‘I will A in C in order to realize or produce E’, where “A” is some type of act, “C” is some type of circumstance, and “E” is some type of end that is to be realized or achieved by doing A in C.[[21]](#footnote-21) As maxims of this form are subjective, and thereby reliant on a posteriori conditions, the first formulation of the CI states that we should only act on those maxims of which we can will that they become universal laws. That is, only if we can still will to act on our maxims if everyone else, when confronted with C, would decide to adopt this same maxim does this maxim become a universal principle. Only if a maxim can be made into a universal principle in this way does it then become moral.

Where the first formulation of the CI highlights the form of moral statements, the second formulation is supposed to add content to these statements. One should always act ‘in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means but always at the same time as an end’.[[22]](#footnote-22) This formulation, which demands a certain fundamental respect for humanity as such, does not prevent us from treating other people as means all together. We may still make use of other persons, but only if in doing so we consider that other person at the same time as an end in themselves. The fundamental property of humanity that warrants the respect that the second formulation expresses is its autonomy, a concept to which I will return in detail in chapter 4. In short, it is because people are capable of applying the CI to their own behaviour, because they have the ability to self-legislate their wills with this universal principle that they are never to be treated merely as the means of other persons.

The third and final formulation of the CI, which, according to Kant, unites the two earlier formulations in a single principle, states that we are to ‘act only so that your will could regard itself as giving universal law through its maxim’.[[23]](#footnote-23) Though this formulation might at first sight seem a lot like the first formulation, its insistence on the will ‘giving universal law’ emphasizes that in addition to our following the universal principle, we are to regard ourselves as the legislators of that universal principle as well. This, again, highlights the importance of our autonomy, of our ability to impose universal laws upon ourselves if we are to act morally.

All three formulations of the CI show its rational, a priori nature. So what is it in Kant that makes him equate morality with rationality? The fundamental reason for this equation of the rational with the moral law then is that morality, if it is to be valid, has to be so universally and unconditionally. We can never say that a certain action is morally good if its being good is dependent on our specific personal situation, or if it is only good in virtue of some end that we are trying to achieve. Only a priori, rational considerations are able of providing us with universal and unconditional conditions, and for this reason we should see morality first and foremost as a rational practice. Though I will hereafter use the terms moral law and rational/universal law interchangeably, it is important to point out here that a law’s being moral is always dependent on its being rational. We are to act only on those maxims that we can at the same time will to become universal laws for instance, not because this is some moral intuition that we have, but because not doing so would imply us to act irrationally. By willing to act on a maxim that we cannot at the same time will to become a universal law, we would be willing inconsistently, that is, we would want ourselves to become an exception to a principle that we want everyone else to obey. It is this rational basis of morality that makes Kant’s moral philosophy so attractive to machine ethicists, since machines are nothing if not rational.

The rational nature of Kantian ethics, in combination with the disadvantages that utilitarianism and virtue ethics pose for machine ethics - as described in section 1.2 - explain why the Kantian position has become dominant within this specific field. This, of course, does not mean that Kant’s moral philosophy is thereby free of critique, or should be accepted as a superior system of moral thought. There are numerous classical and modern forms of critique on the basic principles, methodology and implications of adopting a Kantian perspective on ethics, and these critiques cannot simply be ignored or discarded just because the theory is compatible with the field of machine ethics. It is important to stress here however that it is not the aim of this thesis to provide a philosophical defence of Kant’s position on morality as such, or even to defend adopting a Kantian perspective in machine ethics. The true aim of this thesis, rather, is to conceptualize what it would mean exactly for machines to display moral behaviour from a Kantian point of view. That is, analogous to Alan Turing’s efforts in conceptualizing a Turing machine, I simply aim to make explicit what properties a machine would need in order to become a true Kantian AMA, and propose ways of implementing such properties. This project is relevant, I believe, because as chapter 2 will show, current attempts at Kantian machine ethics fall short in their conceptualization of - or fail to meet the conditions that they have set on - true Kantian moral agency.

# Attempts to conceptualize a Kantian machine, successes and failures

With at least the basic tenets of Kant’s practical philosophy established in chapter one, it is time now to turn to the application of these ideas in the young field of machine ethics. In this chapter, I provide an overview and comparison of the most important Kantian approaches towards machine ethics from the last twenty years. Having done this, I will evaluate to what extent these approaches have been, and consider themselves to be successful in conceptualizing true Kantian AMA’s, and discuss what we might learn from their attempts.

## 2.1 Powers’ three approaches to applying a formalized CI

Perhaps the most well-known work in Kantian Machine ethics is Thomas Powers’ *Prospects for a Kantian machine*, which focuses mainly on the logical systems necessary to translate Kant’s ethics into something which machines might understand.[[24]](#footnote-24) In order to achieve this, Powers begins by pointing out the similarities between human moral reasoning and practical reasoning in machines.[[25]](#footnote-25) Moral reasoning in both humans and machines, Powers argues, is concerned with transforming descriptions of specific factual situations into action statements; I encounter situation A, and formulate an action B that I deem ethically appropriate for A. In addition to this, both humans and machines might learn to incorporate what they have learned in their past moral reasoning in their present considerations, through reflection or the implementation of feedback loops respectively.[[26]](#footnote-26) For humans, the intermediate step that allows us to formulate an appropriate action for each situation consists of the set of normative principles that we adhere to. Note that Powers does not claim humans to be natural Kantians themselves, as the normative principles that we use in our ethical decision making differ from person to person. In order to simulate the normative principles that will allow machines to generate an action-based output from its fact-based input like humans do, Powers suggests that we formalize the first formulation from the categorical imperative that I discussed in chapter one, that is: act only on that maxim through which you can at the same time will that it should become a universal law.[[27]](#footnote-27) When we formalize this first formulation of the CI, Powers claims, we end up with a method for testing maxims (which we have characterized as being subjective determinations of the will, though Powers simplifies this to mean plans for action)[[28]](#footnote-28) on their universalizability. Depending on the results of this test, maxims can then be mapped upon the deontic categories of being forbidden, obligatory or permissible, so that the machine would never allow maxims that do not correspond to the CI to spur it to action. Rather than testing each maxim that comes up in a decision making process to the CI directly, Powers proposes that it would be more efficient to create and test a complete database of all possible maxims m before any decision making actually takes place, so that the machine has the set of forbidden maxims F available for itself at all times. In this way, our machine would only have to check if the maxims m that it encounters in specific ethical situations fall within F to know whether or not it is morally allowed to act on these maxims.

Simply testing maxims to the formalized CI to determine their universalizability as sketched above is Powers’ first approach to a Kantian machine. He immediately rejects this ‘consistency’ approach however as being too ‘thin’ of an interpretation of the CI.[[29]](#footnote-29) By checking to see whether or not a maxim is universalizable, his argument goes, only those contradictions that are inherent to the chosen maxims themselves will end up in the set of forbidden maxims F. This makes the contradictions found by the universalization process trivial , and would allow the machine to classify some maxims that Kant would not consider to meet the categorical imperative as permissible. In order to overcome this problem, Powers’ second approach towards Kantian machine morality focuses on what he calls a background theory of morality B, consisting of the results of earlier ethical deliberations and so called common-sense notions.[[30]](#footnote-30) These common-sense notions or ‘postulates of rationality’[[31]](#footnote-31) are formulated in the *Groundwork* and are meant to capture what it is to reason practically, to provide our machine with rational axioms.[[32]](#footnote-32) By adding a mechanism that checks for inconsistencies with B to the categorical imperative test, this second approach ensures that each maxim that survives the testing procedure is completely rational. As common-sense postulates of reason cannot simply be negated by counterexamples however, the introduction of B requires that we adopt a non-monotonic logic. Non-monotonic, in contrast to monotonic or classical logic, allows for certain logically correct inferences to be superseded or ‘defeated’ by other considerations, so that conclusions may be retracted when the reasoner receives new information. Powers illustrates the need for a non-monotonic logic by appealing to Kant’s famous example of false promises. Making a false promise in times of need is wrong, according to Kant, because universalizing such a rule would make the entire act of promising impossible: no person would ever trust another to promise them anything again, making it impossible to will that making false promises in times of need becomes a universal rule.[[33]](#footnote-33) This sort of reasoning is non-monotonic, according to Powers, because promising is itself defeasible in nature, as promises can be broken on occasion without this causing the  whole institute of promising to collapse. It is only when the rational notion ‘promises are kept’ is defeated more than it is confirmed that this rational notion itself gets discarded, which is not something that monotonic logic manages to capture. One severe problem of non-monotonic systems however is that they fail to meet another formal requirement of machine ethics: semidecidability of a set-membership, or the possibility of determining with certainty whether or not particular statements belong to a set. If the reasoning that we use in checking a maxim’s consistency with B is non-monotonic, then there is no way to make sure that we will find either a positive or negative result for that specific maxim. This is problematic for machine ethics, as allowing for the possibility that our machine might not be able to determine the ethical permissibility of certain maxims is impractical at best, and could have severe ethical consequences in itself. For this reason, Powers rejects his second approach as well.

Powers’ third account is one of coherency, in that it specifies a set of internally consistent maxims that serve as the background to which new maxims should be tested. In this ‘bottom-up’ approach the subject (or machine) builds a theory G consisting of a set of permissible maxims and their consequences M. Rule R-in decides that any new maxim m is allowed in M iff it is consistent with the other m’s already present in M. Rule R-out decides that any previously accepted m will be removed from M if it turns out that this m makes M inconsistent (this is to allow for reflection upon earlier faults in ethical reasoning). How this updating process R-out is supposed to work however remains unclear. A much bigger problem for this approach however is the problem of priority. How is a system to know if a new m is inconsistent with M because of its own moral properties or because some earlier adopted m was mistakenly incorporated in M? The bottom-up constructivist approach ultimately fails, according to Powers, because unlike the other two, it does not incorporate the CI as its ultimate measure of comparison.

Powers’ three approaches to formalizing the categorical imperative represent three ways of interpreting that CI itself, of which only the second comes close to actual Kantianism. If the decidability issues with non-monotonic logic can be resolved, or a way of incorporating the rational background theory that the second approach adds to the first can be found that only uses monotonic logic, this strategy seems like a viable first step towards a Kantian moral machine. What Powers is not aiming at however is the genuine Kantian agency that would be needed for a machine to become moral in itself, to become a fully-fledged AMA.[[34]](#footnote-34) Powers’ hypothetical machines, by testing their optional plans for future action to both the CI and a background theory of rational postulates, will, because of this consistency check, always act *in accordance* with duty. As Powers realizes himself however, there is nothing that would indicate that such machines would act *from* duty as well.[[35]](#footnote-35) They might legislate themselves in the sense that they are the ultimate judge that determines which maxims will and will not generate their actions, but as this process takes the form of an automated consistency check, we would hardly call such machines autonomous. What is self-legislated here is behaviour and not will, as Powers’ hypothetical machines have no ‘will’. In his view, there are only maxims and a method for determining which actions are reasonable and correspond to the CI, but nothing that is remotely capable of ‘grasping’ that CI, of actual moral agency. Even within his limited interpretation of Kant that does not aim at genuine Kantian morality, Powers says that all  three of his approaches ultimately fail.

## 2.2 Ulgen’s Kantian Machines

Where Powers largely ignores the intricacies of Kant’s philosophical work and focuses his efforts mainly on conceptualizing a computable categorical imperative, Ozlem Ulgen’s *Kantian Ethics in the Age of Artificial Intelligence and Robotics* engages with Kant on a much deeper level.[[36]](#footnote-36) In particular, and very much relevant to our purposes here, Ulgen focuses on what would be needed for a potential Kantian AMA to become autonomous, that is, what a machine would need to rationally choose to follow the CI. This marks a fundamental difference in the goals that Powers and Ulgen strive towards. Where Powers is aiming for an implicit moral machine in Moor’s sense, Ulgen is looking at what it would take to create explicit or even full moral agency.

Like Powers, Ulgen takes the first formulation of the Categorical Imperative as her starting point, and sees this CI as a test that is capable of determining which maxims should be used to form universalizable action-statements.[[37]](#footnote-37) Unlike Powers however, Ulgen emphasizes that the only maxims (or in her jargon, rules) that the CI allows for are those maxims that are derived from rational thinking, and that it is a requirement for moral action that the CI is followed autonomously by a rational and free being. Having a machine that follows an algorithm that simulates the CI simply won’t do for Ulgen, since, as I pointed out above, this machine would act only in accordance with duty. In order to focus on this larger goal of artificial moral agency then, Ulgen further examines the Kantian concepts of autonomy and rationality in relation to the field of machine ethics.

Autonomy and rationality, as we have seen in chapter one, are two fundamentally related concepts in Kant’s thinking. An autonomous will is a will that self-legislates itself, but this self-legislation can only really be called autonomous as long as it directs itself according to rational principles. That is, a will that legislates itself to follow the subjects’ every desire can never be called autonomous, because by doing so it would hand over the authority to direct itself to those subjective desires. The rational autonomous will makes a choice to apply the categorical imperative rather than to let itself be directed by the subject’s contingent desires and inclinations, precisely because doing so is a rational choice. Rationality, as Ulgen defines it, is that faculty that provides us with the principles of cognition a priori; that is, it is concerned only with those considerations that we can know before having any subjective experiences.[[38]](#footnote-38) It is for this reason that the rational will chooses the CI as its motivator over the subject’s desires: desires are empirical and a posteriori, and can therefore never be the object of pure rational deliberation. It is this capacity for acting autonomously, as we have seen in chapter one, that provides rational beings with the dignity that the second formulation of the CI calls for. This formulation, (act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means but always at the same time as an end)[[39]](#footnote-39) firmly establishes the intrinsic value of humanity, which is based on its capacity for acting autonomously. We shall return to what it means to act reasonably exactly in chapter 3. Suffice it to say here that for Ulgen, Kantian moral machines need to be able to rationally choose to act in accordance with the CI if they are to be moral at all. Having set out these core principles of Kantian moral philosophy (the CI, rationality and autonomy), Ulgen considers the extent to which each of these principles can be applied to machine ethics. Ulgen identifies problems with the practical implications of all three of these aspects in machine ethics, leading her to conclude that the true Kantian artificial moral agent is impossible, and that the ultimate moral responsibility for any Kantian machine would remain in human hands.[[40]](#footnote-40)

Properly applying the CI, according to Ulgen, is impossible for machines because they lack the ‘human rational thinking capacity’[[41]](#footnote-41) that is needed to determine whether or not a maxim is universalizable. Having machines generate their own rules would lead to situations in which the rules that are generated are not in accordance with the CI, and for this reason the only Kantian possibility would be to have humans program correct (CI-corresponding) rules into machines beforehand.[[42]](#footnote-42)  Providing a machine with autonomy of the will runs into its own set of problems, according to Ulgen, because machines lack the understanding of concepts such as ‘dignity’ or ‘rule-making capacity’ and are therefore unable to value and apply these concepts.[[43]](#footnote-43) In addition to this, machines lack a will, or the ‘self-determining capacity that can make choices between varying degrees of right and wrong’.[[44]](#footnote-44) Machines can follow rules, but they do not decide to follow these rules; they do not *choose* their rule-following behaviour freely from a set of possible actions.

Ulgen considers the second formulation of the CI to treat people never solely as means towards separate ends only in the context of autonomous weapons systems.[[45]](#footnote-45) Within this military application of intelligent machines, Ulgen concludes, human dignity is always sacrificed in order to achieve certain utilitarian goals. Autonomous weapons systems are designed to kill certain military targets, and in doing so treat human life as a relative, rather than an absolute or objective end. Since human beings are always to be treated as ends in themselves, autonomous weapon systems violate the second formulation of the CI. The prospects for machines respecting human autonomy in civilian contexts remains unconsidered by Ulgen, leaving open the question of whether or not Kant’s second formulation of the CI can be fulfilled by machines period.

Ulgen clearly sets up three demands that a Kantian machine needs to fulfil in order to be considered an AMA. In considering the extent to which these demands could be met, she concludes that this ultimate goal of explicit Kantian machine agency is impossible, because Kant’s ethics are too human-centric for machines to properly comprehend and apply. Her specific arguments for why machines are unable to apply the CI, display rational autonomy or respect human dignity however are quite limited. Without much consideration or argumentation Ulgen concludes that machines lack the capacities needed for meeting these demands, and comes to her negative conclusion about the Kantian AMA.

## 2.3 Other approaches and considerations from machine ethicists

Though Powers and Ulgen provide the most extensive theories of Kantian machine ethics, Kant’s deontological ethics and the CI are mentioned in a number of other works on machine ethics from a broader perspective. This section discusses some valuable contributions on the subject by Wallach & Allen (2010) and Anthony Beavers (2009).

Wallach and Allen, in their excellent *Moral machines, Teaching Robots Right from Wrong*, consider the categorical imperative in their chapter on top-down deontological approaches towards machine ethics.[[46]](#footnote-46) This CI, as in Powers and Ulgen, is interpreted by Wallach and Allen to serve as a formal thinking tool that machines could use to evaluate ‘behaviour-guiding’ maxims.[[47]](#footnote-47) Like Powers, Wallach and Allen focus their efforts only on the possibility of implementing the CI in machine decision making processes, and therefore do not aim for full or even explicit moral agency. Like Ulgen however, they formulate a concrete set of conditions that machines would have to meet in order to achieve this goal.[[48]](#footnote-48) In order to successfully engage with the CI, according to Wallach and Allen, machines would need the ability to formulate explicit goals (1) and propose plans (or maxims) that aim to achieve those goals (2). In addition to this, our machines need to be able to universalize their maxims in what Wallach and Allen envisage to be a simulation program (3), and determine under what circumstances this universalization is to be evaluated as being successful (4). As this simulation would involve predicting and evaluating the social implications of universalizing said plan, our machine needs a firm grasp of machine and human psychology and reasoning (5). Though Wallach and Allen do not go into the specifics of what it would mean to implement these five faculties in machines, they seem at least tentatively optimistic about the feasibility of such a project.[[49]](#footnote-49) The reason that they abstain from further developing the Kantian approach to machine ethics then is not a technical, but rather a moral one. Top-down approaches in general, and thereby Kantianism as well, are simply too rigid in Wallach and Allen’s opinion to adequately deal with the ambiguous nature of ethics. Even if our Kantian machine meets all five conditions and consistently acts in accordance with the CI, there will always be particular situations in which the human observer will evaluate the machine’s behaviour as ethically wrong, because moral reasoning in humans simply does not yield the cold, logical results that a Kantian machine would produce.[[50]](#footnote-50) As this argument is based on assumptions about moral reasoning in humans, it would be beyond the scope of this thesis to engage with it directly. Ignoring this ethical argument then it seems that Wallach and Allen provide us with a concrete conceptualization of what it would take to create an implicit Kantian machine.

Returning to the field of full moral agency, Ryan Tonkens’ *A Challenge for Machine Ethics* argues that the creation of true Kantian moral agency in machines, if even possible, would violate the categorical imperative in and of itself. Tonkens argument starts off with a well-considered overview of what it means to be a Kantian moral agent, culminating in the definition that Kantian moral agents are to be and act both free and rational.[[51]](#footnote-51) Because machines, according to Tonkens, are completely determined by their programming, they do not have a free will and therefore lack one of the conditions necessary for Kantian agency.[[52]](#footnote-52) If a machine is ever to become a true moral agent, it’s creators (we) would have to endow it with freedom, just as God, our creator in Kant’s thinking, has endowed humans with freedom. In addition to this, Tonkens holds that creating a Kantian AMA would violate the second formulation of the CI, because it would treat those AMA’s only as means, and not as ends in themselves.[[53]](#footnote-53) Though we would demand, by adopting Kant’s deontological framework, that our AMA’s treat every other agent as an end in themselves, we would not apply this demand to those AMA’s ourselves. Because of this inconsistency in our wish for an AMA, the maxim to create Kantian AMA’s is itself not universalizable, which means that the first formulation of the CI is violated as well.[[54]](#footnote-54) It remains unclear from Tonkens text however why the Kantian AMA, if it is to lack genuine Kantian agency,  would be deserving of the dignity that the second formulation assumes. If Tonkens is right, then AMA’s are no true moral agents due to their lack of freedom. Since the second formulation of the CI only applies to moral agents then, there would be no reason for us at all to treat these AMA’s as ends in themselves.

In parallel to, but from a completely different argument than Tonkens, Anthony F. Beavers has argued that it is immoral to create a genuine Kantian AMA, because such AMA’s would need to have the freedom to not act in a moral way.[[55]](#footnote-55) It is a crucial fact about Kantian agency, in Beavers’ reading of Kant, that a moral agent is suspended between inclination and duty, that is, that their choice to act morally always entails the possibility of following one’s inclinations instead. Building a machine that is able to act immorally however, to sometimes ignore duty and follow its inclinations instead would be, according to Beavers, obviously immoral in and of itself.[[56]](#footnote-56) For this reason Beavers argues that we should, in our attempts at a Kantian machine, forfeit the demand for agency by omitting the possibility for immoral behaviour. Such machines might not be true AMA’s, Beavers admits, but they are to be preferred over machines with the potentiality to act immorally.[[57]](#footnote-57)

## 2.4 The challenges for a Kantian machine

As we have seen in this chapter, contemporary approaches to Kantian machine ethics can be divided into two categories. On the one hand we find those who, like Powers Wallach and Allen, seek for implicit moral agency by merely conceptualizing a computational categorical imperative. On the other we find Ulgen, Tonkens and Beavers, who aim for full moral agency in machines and engage with questions about autonomy, freedom and the will. In considering these approaches, we have come across a range of potential problems for Kantian moral agents. These problems or challenges for a Kantian moral machine can, in general, be divided in three categories.

First of all, we have come across challenges that are related to the faculty of practical reason and the proper implementation of the CI. In this category we find Powers comments about the nonmonotonic nature of practical reasoning and Ulgen’s doubts on the possibility of machines being able to universalize maxims and have a will. I deal with these objections in chapter three, in which I set out a preliminary account of machine rationality and a computational CI based on the account of Wallach and Allen described above.

The second domain of challenges to a Kantian AMA concerns questions about freedom and autonomy. As we have seen a number of times now, it is crucial for any Kantian moral agent that the subject self-legislates its own will to follow only those demands that it receives from reason, as opposed to those that follow from their inclinations. The ability for machines to do this is called into question by Ulgen and Tonkens, and will be considered in chapter four. Closely related but conceptually distinct from this demand for autonomy are questions about free will in machines, as put forward by Tonkens and Beavers among others. Chapter four will consider the relation between autonomy and freedom in Kant, and proposes a compatibilist interpretation of his work based on and Pauline Kleingeld’s interpretation of Kant’s freedom of the will.[[58]](#footnote-58) Reading Kant in this compatibilist manner, I will argue, allows us to conceptualize both the autonomy and freedom that a Kantian agent needs within the limits of programmable machines.

Finally, chapter five will address questions about the ethical permissibility and general desirability of creating Kantian AMA’s. As Tonkens has pointed out, it seems that success in creating an actual agent would bring with it the duty to treat those agents as ends in themselves. In addition to this, we’ve seen Beavers argue that it is inherently immoral to create anything with the capacity to act immoral itself. I will compare the basic design for a Kantian AMA developed in chapters two and three with the AMA’s used in Beavers’ and Tonkens’ arguments, and assess the Kantian implications for creating such an AMA.

# The categorical imperative and the reasonable will

## 3.1 The CI: pure analysis vs simulation

As we have seen in the previous chapter, beginning a theory of Kantian machine ethics by conceptualizing the ways in which machines could use the first two formulations of the CI to check its possible courses for action has become common practice. As we have also seen however, being able to apply the CI, though necessary, is not sufficient for actions or agents to be considered moral.[[59]](#footnote-59) A CI-following machine, as described by Powers and Wallach & Allen in the previous chapter is only implicitly moral, and therefore does not qualify for the strict conditions Kant sets on agency. Still, for an AMA to achieve such full Kantian agency, it would need to be at least capable of correctly applying the CI and to reason a priori about the results of universalizing certain maxims. This chapter is therefore concerned with simulating only these implicit aspects of a Kantian AMA, so that in chapter 4 this preliminary account may be adapted into something truly worthy of the name of a Kantian AMA. So far we have come across two viable approaches to a computable CI. First, we have seen Powers’ second attempt at a formalized CI, which took the form of a computer performing a nonmonotonic logical analysis on linguistic maxims. Second, we’ve seen some careful speculation by Wallach and Allen, who conceived of a computable CI in terms of computers running concrete simulations, rather than logical analyses. In what follows, I will argue for and further develop this second approach, based on Powers’ concerns about the lack of semidecidability in nonmonotonic logic and it’s closer resemblance to the way Kant envisages humans using the CI. First, let us recall and explicate the differences between Powers’ and Wallach and Allen’s approaches, so that we may critically compare them to the human application of the CI as envisaged by Kant.

For Powers, applying the CI is a fundamentally linguistic procedure in which maxims, in the form of propositions, are universalized and checked for internal inconsistencies.[[60]](#footnote-60) Since this process only filters out internal, trivial inconsistencies however, Powers adds the set of common sense reasonable propositions B, against which the universalized maxims must be tested as well (see chapter two). Because of the nonmonotonic nature of moral reasoning however, testing a maxim’s consistency with B is not semidecidable, and the whole approach is rejected as a viable system for machine ethics by Powers himself.[[61]](#footnote-61)

Wallach and Allen’s approach, though much less further developed, sketches a completely different picture of how machine’s might apply the CI in their ethical deliberation: *‘... an AMA would need an explicit and fully stated principle of practical reason consisting of three elements: a goal, a means or course of action by which the agent proposes to achieve that goal, and a statement of the circumstances under which acting in that way will achieve the goal in question. Given these three elements, a very powerful computing device might be able to run an analysis or a simulation model to determine whether its goal would be blocked if all other agents were to operate with the same maxim’*.[[62]](#footnote-62) After having set a maxim (a means or course of action by which the agent proposes to achieve a certain goal), Wallach and Allen’s machine does not simply perform a logical transformation on this maxim, but engages in a simulation that models a world in which every actor has adopted this maxim universally.  Having done this, the machine checks the extent to which it, within this simulated world, would still be able to perform the maxim now adopted by everyone else (a statement of the circumstances under which acting in this way will achieve the goal in question). In order to realistically simulate the social impacts of universalizing a maxim, the machine needs a database of psychological knowledge and a highly developed understanding of semantics.

A first important argument for adopting the simulation, rather than the logical analysis approach, is that it better captures how Kant himself envisages the role and function of the CI. As we have seen in chapter one, knowing whether or not a maxim is universalizable in and of itself tells us nothing about the ethical permissibility of adopting that maxim. What is relevant for Kant, as Wallach and Allen seem to realize, is the extent to which an actor would be able to successfully use a certain maxim if that maxim was adopted universally.  If the universal adoption of a maxim blocks actors from successfully adopting and executing that maxim themselves, the maxim entails what Onora O’Neill has called a conceptual inconsistency, which disqualifies this maxim from becoming a universal law.[[63]](#footnote-63) Conceptual inconsistency in a maxim means that the successful implementation of that maxim in a world in which it is universally held cannot be thought consistently, that it is not possible to a priori formulate this maxim as a natural law.[[64]](#footnote-64) We see this sort of inconsistency clearly in an example already discussed in chapter two, in which a person in severe distress decides to lend money under the false promise that they will pay it back at a later time. Universalizing the maxim to give false promises in times of need, as we have seen, would undermine the institution of promising itself, making it impossible for us to consistently think of making false promises in a world in which everyone does so; the maxim is conceptually inconsistent.[[65]](#footnote-65) In addition to these conceptual inconsistencies, O’Neill emphasizes that for Kant, maxims may turn out to entail volitional inconsistencies as well. If, after universalization, I am still theoretically capable of acting on my maxim, but I can no longer rationally will to do so, a volitional inconsistency has been found and this maxim is to be discarded. Kant himself emphasizes the importance of these volitional inconsistencies (a term he does not use himself) in one of the famous examples of duty in the second section of the *Groundwork*. Suppose that a person who is quite well off, and who sees others around him struggle in their day to day needs decides to adopt the maxim: ‘I will not take anything from him or even envy him; only I do not want to contribute to his welfare or to his assistance in distress’.[[66]](#footnote-66) Now a person would be quite able, Kant proclaims, to act upon this maxim in a world in which every person who is quite well off and encounters others who are not decides to adopt this maxim; the maxim is not conceptually inconsistent. There is nothing inherent to a world in which every rich person decides not to help others that would limit my individual capacity to act this way as well. It is not possible however, Kant maintains, to will that this maxim gets adopted universally without exception. There is always a possibility that the person who adopts this maxim will at some point in their life want to receive the sympathetic participation from others themselves. Since this wish would contradict with the wish to universalize the maxim above, this maxim entails an inconsistency in the will, or to speak with O’Neill, a volitional inconsistency.[[67]](#footnote-67)

Why do I claim that this explanation of the way in which the first formulation of the CI works in Kant serves as an argument for adopting the simulation, rather than the logical analysis approach? By actually simulating the agent’s attempt to act on its maxim after that maxim has been made a universal law, I propose, our Kantian machine engages in a process that is much more similar to the thinking described in the examples above than the logical analysis proposed by Powers. The subject (or machine) situates themselves in the world in which their maxim is made universal law, and proceeds to check whether the universal law prevents them from acting on their maxim. Performing the logical transformation ‘universalize’ on a maxim, and checking for inconsistencies between this universalized maxim and a set of common sense background notions seemingly ignores the condition that the subject must themselves be able to think of and will their maxim to become universal. It checks for rational inconsistencies in universalized maxims, but fails to capture the agent’s own relation to these inconsistencies. Where a simulation would show that acting on a maxim that fails the CI would entail a subject making an exception out of themselves by exempting them from rules that they subject others to, a logical analysis simply shows that some maxims themselves are irrational, and forbids the subject from acting on those maxims.

A second advantage of the simulation over the analysis approach can be found in Powers’ own claim about the undecidability of nonmonotonic logic. Because we want our machines to be at least semidecidable according to Powers - and I agree with him on this point - a nonmonotonic moral machine is undesirable, which means that Powers’ analysis approach incorporating a background theory of common sense notions is untenable. By modelling the social implications of universalizing maxims however, I claim, the simulation approach is not in need of incorporating nonmonotonic logic at all, and therefore runs free of the problem of undecidability. In order to show this, let us reflect upon the reasons why Powers introduced nonmonotonic logic in the first place, and see how the simulation approach would deal with these reasons. Promising, as we have seen in the previous chapter, is defeasible in nature, which means that there exists a certain tipping point at which making a false promise becomes an irrational thing to do. To incorporate this defeasibility in a logical analysis, Powers claims, nonmonotonicity must be introduced, so that we can make sense of the fact that false promises only  become irrational once a critical amount of false promises has been made. If we suppose, as Wallach and Allen do, that our Kantian AMA has sufficient knowledge of human reasoning and psychology to reliably and realistically simulate social behaviour in situations where everyone adopts a certain maxim universally, the need for nonmonotonic logic disappears. If the AMA can correctly predict at which point humans would consider promising to have become an irrational thing to do, there is no need for the AMA to engage in nonmonotonic reasoning itself. It simply predicts the outcome of humans engaging in nonmonotonic reasoning, and uses this prediction to complete its simulation.

  One important caveat to simply copying Wallach and Allen’s simulation approach however is the fact that this approach only seems to incorporate a ban on conceptual inconsistencies, and seemingly ignores the equally important volitional inconsistencies described above. At first sight, it seems that in order to incorporate the volitional inconsistencies in Wallach and Allen’s framework, we could simply expand their third aspect of practical reason (a statement of the conditions under which acting in a certain way will achieve the set goal). Where their original statement of these conditions reads: ‘whether its goal will be blocked by other actors acting under the same maxim’, we would add to this the condition that achieving the set goal under the formulated maxim could never entail a contradiction in the machine’s will. As Wallach and Allen’s AMA does not have the faculty of  ‘will’ under its present assumptions however, it remains unclear what this extra condition would mean exactly. We have now stumbled upon a distinct, but closely related problem with Kantian AMA’s, already pointed out by Ulgen in chapter two.[[68]](#footnote-68) Machines, or at least the machines that we have at this point in time do not have what humans understand as a ‘will’, which, in addition to the problem of volitional inconsistencies makes it difficult to understand how such machines could ever be called autonomous in the Kantian sense.

## 3.2 Having a will, perfect and imperfect rationality

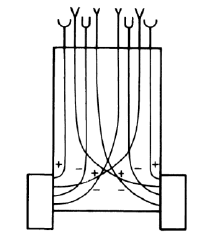
What does it mean to have a will, and can we conceptualize what it would mean for a machine to have such a thing? Ulgen herself tentatively defined the faculty of will missing in machines as a ‘self-determining capacity that can make choices between varying degrees of right and wrong’[[69]](#footnote-69), but this definition is not necessarily based on Kant’s conception of the will and lacks proper argumentation. For Kant, having a will is a property reserved for rational beings only, in the sense that only they are capable of acting in accordance with principles, with representations of laws. ‘Acting in accordance with a law’ then implies that these rational beings are capable of deriving concrete actions from abstract law, and it is this capacity that Kant defines as ‘will’ or practical reason.[[70]](#footnote-70) Objects of nature, such as animals or physical objects, are ‘governed by natural law’ in the sense that those laws determine their behaviour. An animal makes representations of the desires that they have, and these representations exert a certain pull on that animal. When one of these pulls becomes stronger than the others, when that desire is decisively preferred, the animal ‘chooses’ to act on this desire with what Kant calls an arbitrium brutum or animal will.[[71]](#footnote-71) Beings with a proper will or practical reason however have the capacity to reflect on their desires, to conceptualize the representations of their desires and to think and reason about them.[[72]](#footnote-72) It is this conceptualization that allows rational beings to act in accordance with the law, to justify their particular desires with reference to universal principles and select on which desire to act with arbitrium liberum, the free power of choice.[[73]](#footnote-73)

Beings endowed with practical reason, for Kant, come in two distinct forms. On the one hand we have beings whose reason determines the will infallibly, that is, beings who always act in accordance with the rational law that is the CI. On the other hand there are those beings whose reason is not sufficient to determine their wills, that is, whose wills are susceptible to other laws than those of pure reason.[[74]](#footnote-74) For a purely rational being, reason manages to determine the will infallibly. For such a being, what is subjectively necessary - that derivation from a law that will ultimately motivate the being’s action - by definition always overlaps with what is objectively necessary, that is, with those actions that are universally valid for all rational beings. The purely rational being, in other words, always wills those things that correspond to duty, always chooses what reason recognizes as the good because it only has this reason to provide it with laws to act in accordance with.[[75]](#footnote-75) For beings whose reason only fallibly determines their will, in contrast, the objectively necessary is subjectively contingent, in the sense that the law-derivation that ultimately motivates their actions is not necessarily the same as what the objective law would prescribe them to do. This is the case because these beings, in addition to reason, experience other motivational impulses, such as feeling and inclination, as well.[[76]](#footnote-76) Fallible wills certainly have the capacity to heed the call of reason, but they are not necessarily determined to do so, as they can also allow their other motivational impulses, which are given to it by the laws of nature, to govern its will.

If we want to know what it means for an AMA to have a will, we first have to determine what sorts of will we would want these AMA’s to possess. If AMA’s are to have an infallible will, we would only have to give them the capacity to derive concrete actions from the reasonable law, that is, to apply the CI correctly. Determining volitional inconsistencies then, the problem that gave rise to this discussion of the will in the first place, only seems a matter of comparing current maxims with the complete set of maxims that the CI allows for. As the CI should never yield conflicting objective laws due to its rational and objective nature however, infallible wills would never encounter volitional inconsistencies at all. As a Kantian machine that possesses an infallible will would always act in accordance with duty, and seems relatively easy to create - in that it needs no ‘strange’ human faculties, and could simply apply the CI simulation program described above - this machine seems like the perfect candidate for the creation of a Kantian AMA. There exists a general agreement among Kant scholars however that the behaviour of Kant’s perfect rational beings (angels and God), though corresponding to the dictates of the objective rational law, is amoral.[[77]](#footnote-77) Precisely because such perfectly rational or divine wills do not have any other option than following reason, and are therefore determined to act in accordance with the CI, the term ‘ought’ is out of place here, meaning that these beings are unable to act from duty.[[78]](#footnote-78) In order to be truly capable of morality the way humans are, beings need to have an infallible will, to be affected both rationally and sensibly. This illuminates Beavers’ earlier statement about being suspended between inclination and duty; only those beings that experience the pull of both their natural senses and their rational duty are able of overcoming the former in favour of the latter, and can thereby be said to act morally.

Since the larger aim of this thesis is to conceptualize what it would mean for a machine to display full Kantian agency, we may reject the infallible will at this point and continue on with the fallible reasonable will. If our AMA is to have such an infallible will however, our problem is complicated further. Such an AMA would need to be susceptible to motivational impulses other than reason, that is, it would need to have inclinations towards certain behaviour that do not necessarily correspond to duty. In addition to this, it needs some faculty that allows it to decide what is to determine its will, that is, what law it is to derive its actions from, the rational law of the CI or the natural law that determines its inclinations. Once a machine has these two faculties, we may conceptualize what it would mean for it to encounter volitional inconsistencies.

One promising method of incorporating inclinations in machine intelligences has been proposed by Anthony Beavers, who discusses the application of so-called Braitenberg vehicles in relation to Kantian AMA’s.[[79]](#footnote-79) These vehicles were first introduced by Valentino Braitenberg in 1984 as a set of thought experiments that aimed to mimic complex animal behaviour in mechanic vehicles by endowing those vehicles with several sensors, and allowing them to move around autonomously based on the information received through these sensors.[[80]](#footnote-80) Though Beavers only briefly mentions the potential benefits of incorporating elements from Braitenberg’s machine into a Kantian AMA, I would like to spend some time here discussing the specifics of Braitenberg vehicles and the ways in which they could be used to supply a Kantian AMA with inclinations.

 Braitenberg’s first and simplest vehicle only has one sensor to measure its surrounding temperature, and one motor that is wired so that it’s propelling force increases as the vehicle encounters higher temperatures and decreases as temperatures drop.[[81]](#footnote-81) In his subsequent thought  experiments, Braitenberg increases the amount and types of sensors of his vehicles, and shows how different wirings between those sensors and the motors that propel them will result in different types of complex movement or behaviour. His final vehicle 3c, (see figure 1) now endowed with four pairs of sensors that measure different qualities, two motors in order to control its direction of movement and two sets of inhibitory and excitatory connectors that allow it to slow down once it reaches its target or speed up when it moves too far away from it displays some very interesting behaviour. *‘It dislikes high temperatures, moves away from hot places, and at the same time seems to dislike light bulbs with even greater passion, since it moves towards them and destroys them. On the other hand it definitely seems to prefer a well-oxygenated environment and one containing any organic molecules, since it spends much of its time in such places. But it is in the habit of moving elsewhere when the supply of organic matter or (especially) oxygen is low. You cannot help admitting that vehicle 3c has a system of values*’.[[82]](#footnote-82) Though Braitenberg admits that his vehicles do not possess cognitive states or proper knowledge about their environment, it seems clear that these experimental vehicles do have inclinations or dispositions towards certain types of behaviour. These inclinations might be incorporated in a hierarchical structure, that allows the stronger inclinations to trump their weaker counterparts (what Braitenberg calls a ‘system of values’). This vehicle seems to bear resemblance to how Kant envisaged animals to choose between their different inclinations: it ‘chooses’ to act on its strongest pull in a determinist fashion, with a faculty that we might describe as an arbitrium brutum. There still exists one big difference between the Braitenberg machine and the Kantian animals described in the beginning of this section however that prevents us from simply incorporating Braitenberg mechanics in our AMA and calling it a day. Where Kant’s animals may choose to act on a certain inclination based on its mental representation of this inclination, Braitenberg’s vehicles act in a completely mechanistic fashion that bears a closer resemblance to Descartes’ thinking about animals than it does to Kant’s.[[83]](#footnote-83) Without a representation of their inclinations, it seems unclear how partially rational beings could translate their inclinations into maxims, and reason about them in the way that humans do according to Kant. This difficulty, however, does not seem conceptually unsurmountable. Suppose that we construct a Braitenberg vehicle resembling 3c, and provide it with a new component that measures the way in which its motor interprets the sensory data it receives. As this motor must have some way to determine its output O in relation to its input I, our new component could send a representation of this process to a central processing unit. Such a representation could take the form of a simple set of propositions, consisting of the general algorithm that the motor uses to determine its output, the particular input I at a specific time and the output O corresponding to that sensory input. The central unit, which does not exist in Braitenberg’s original vehicles, now contains a representation of the vehicle’s inclination to act in a certain way. We might even further hypothesize a rewiring of our vehicle that would ensure that it is not the actual algorithm inside the motor, but the representation of this algorithm inside the central processing unit that ultimately spurs the machine to movement. We now have a machine that contains a central processing unit in which representations of different inclinations are selected to spur that machine into action based on the relative strength of their ‘pull’, a system very similar to the Kantian animal as described above.

Now that we have conceptualized both a Kantian angel (the infallibly rational CI simulation machine) and a Kantian animal (our adapted Braitenberg machine), let us consider to what extent we can combine these two extremes in a being that has both of these capacities, a being that is affected, as humans are, both rationally and sensibly.

## 3.3 Implicit Kantian agency in machines

With what we have said so far, we may give a first attempt at conceptualizing a machine that has a fallible rational will, and is able to successfully apply the categorical imperative in a similar manner to the way humans do so. First, we need our machine to be able to set goals, and formulate plans to achieve these goals, so that it has Kantian maxims that it might insert into the CI. The Braitenberg-like vehicle that I discussed in the previous section provides one way to do this, by reformulating the representations of its inclinations into maxims. If our machine has the inclination to move towards an observed location that is, for instance, particularly warm, then its central processing unit could reformulate the representation of this inclination into a maxim, in which ‘move towards this warm location’ is the goal, and ‘engage the motors in a way that moves me towards that location’ is the proposed plan to achieve this goal. Second, we need our machine to be capable of universalizing its maxims and check to what extent this universalization allows it to successfully think and will to act on this maxim. For this, I have invoked Wallach and Allen’s simulation approach, which models the hypothetical situation in which a maxim is universalized and proceeds to simulate the extent to which the machine would itself be able to act on its maxim. In order to have this simulation program check for volitional inconsistencies as well, I propose a mechanism that allows the simulation program to check for inconsistencies between the maxim it is currently operating on and the set of maxims constructed in the CPU of its Braitenberg structure. If it turns out that it would be possible to act on a certain maxim if that maxim were to be universalized, but that this maxim contradicts another maxim that the machine is likely to ‘will’ at a later point or already wills, then this maxim will be discarded. Note that in order to say that the machine ‘wills’ this other maxim, it too must pass the simulation test.

What we have so far is a machine that is capable of formulating maxims and checking those maxims on their capacity for becoming universal law. Now in order to fulfil the condition that our AMA is to have reason only fallibly determine its will, we need to introduce some law that can serve as an alternative to the CI simulation in evaluating the extent to which maxims are to spur it to action. For this, we can simply use the ‘law’ that our Kantian animal used: act on that representation of an inclination that has the strongest pull. If we call this law the law of nature, and see the CI simulation program as our law of reason, our machine lives, as humans do, in both a world of sense and a world of reason. It experiences the motivational pull of natural inclinations and has the ability to act on the universal law of reason. The only thing that our machine needs now to become a proper Kantian agent is the ability to autonomously choose the rational over the sensible, to freely act from duty in the face of the temptations of its inclinations.

# Freedom and autonomy

This chapter is concerned with questions about autonomy and freedom. We have run into two distinct problems that machine ethicists have established for a Kantian machine concerning these concepts in chapter two, and I will attempt to resolve these issues in this chapter. First, there is the problem of autonomy. As Ulgen pointed out, there exists a legitimate concern that machines lack the understanding necessary to legislate their own will. Machines may be able to follow rules, but they cannot decide to follow certain rules while not following certain other rules, leaving them incapable of acting autonomously. Second, there is the problem of free will. Ryan Tonkens for instance has argued that because machines are always determined to act in a certain way, they simply lack the freedom that would be necessary for them to be true moral agents. Similarly, Beavers has made the point that in order for machines to be moral, they need to be free to act immorally as well. In order to make sense of these questions, I will first clarify the two related but distinct concepts that these questions refer to: rational autonomy and freedom of the will.

## 4.1 Kantian autonomy

Perhaps the most important concept in Kant’s practical philosophy, closely tied with the categorical imperative, must be that of autonomy. As we have seen several times now, autonomy, for Kant, is a type of positive freedom to self-legislate one’s will. Concretely, autonomy is the ability to prescribe to one self the universal, purely reasonable law, which alone may determine the will necessarily.[[84]](#footnote-84) As this law of pure practical reason is the same as the CI, it is this faculty of autonomy that we need for the machine we left at the end of chapter three to become able of acting morally. This machine, as the reader remembers, had representations of its ‘natural’ inclinations and was able to formulate maxims for the acts it was inclined to. By engaging in the simulation program we adapted from Wallach and Allen and checking for inconsistencies with other inclinations capable of universalization under this simulation program, this machine was able to apply the universal law of practical reason. What this machine still lacks at this point however is exactly the autonomy Kant describes. It might have a mechanism for applying the moral law, but it is not yet able to prescribe this law upon itself, to legislate its will with it. Without such a faculty, the hypothetical machine of chapter three remains amoral: it has the ability to apply the CI, but seeing as we have not yet constructed a mechanism for it actually doing so it will simply act on its inclinations (that is, act on the natural law) and leave its faculty of practical reason unused.

One important comment to make here is that for a being to be autonomous for Kant, it is not necessary that this being creates the moral law that it uses for self-legislation all by themselves.[[85]](#footnote-85) It is not a problem therefore that we have implemented a simulation program in our machine that expresses the universal law beforehand, as long as it is the machine, and not us, that legislates themselves to actually follow this law. In order to illuminate this aspect of autonomy, let us consider the source of Kant’s concept of autonomy and what it means exactly. As several authors have pointed out, Kant’s autonomy of the will is closely related to the idea of political autonomy as espoused by liberal thinkers of the 18th century, most notably Jean-Jacques Rousseau.[[86]](#footnote-86) For Rousseau, despotism is characterized by the condition that the law that people have to follow is a law that has its origin ‘outside’ those people, that this law is laid upon them by some external sovereign. A free state, by contrast, is one in which citizens are only bound by those laws that they enact upon themselves, laws which are upheld ‘from the inside’. What is important here is not that every citizen takes part in writing these laws, but rather that they are the ones who impose it on themselves. Kant’s idea of the autonomous will works according to this same principle, where a will that follows laws that are imposed on it from outside itself is heteronomous, and only the will that imposes the law upon itself can be called autonomous.[[87]](#footnote-87) So what does it mean to impose a law upon oneself, and how does Kant imagine humans doing this?

As Paul Guyer has pointed out, autonomy is not the same as complete autarky, i.e. rational self-determination by the complete elimination of all of our inclinations.[[88]](#footnote-88) The autonomous will, rather, overcomes its inclinations and acts in spite of them with reference to the validity of the universal moral law. What motivates this overcoming of the natural inclinations then is a fundamental respect for the moral law, a realization by the subject itself that acting on this law is the right thing to do. This is exactly the aspect of autonomy that Ulgen took issue with. Since machines are not able, according to her, to understand what it means to respect the validity or value of a law, machines will never be capable of autonomy.[[89]](#footnote-89) Now since we are already working within a framework in which machines can have inclinations towards certain behaviour, installing a strong inclination towards acting in accordance with the rational law may seem like a simple step at this point. It is important to highlight however that acting out of respect for a law is not the same as being inclined to follow that law. If we were to somehow install a Braitenberg-like inclination to engage in the simulation program in our machines, and made this inclination stronger than any of our machines other inclinations, our AMA would still only act in accordance with the moral law because it is inclined to do so. We might compare such a machine with Kant’s example of people who take an inner gratification in spreading joy to others, who have an inclination to act in accordance with the moral law.[[90]](#footnote-90) The acts of these people, though worthy of honour, praise and encouragement, nevertheless lack all moral content, because they are done out of inclination rather than duty. A will is autonomous then not because it wants to or is inclined to impose the rational law upon itself, but because it recognizes its validity and therefore respects it; because it acts from duty alone.

If duty is the ultimate and only motivator for prescribing the moral law upon oneself, it is this duty that our machine from chapter three needs in order to bring itself to apply its simulation program to the maxims that it forms out of its inclinations. But what exactly is this duty, and to what extent does this concept lend itself for an application in machine ethics? The first definition that Kant provides in his *Groundwork* is that duty is the necessity of an action from respect for the law.[[91]](#footnote-91) This at least clarifies the relation between duty and the concept of respect for a law; if an action is deemed necessary by a person’s respect for a law, then this action is done from duty. Having respect for the (rational) law is thus a necessary condition for duty, and thereby for autonomy. We are still left in the dark about what it means to respect a law however, as we have only established so far that this respect must be something other than an inclination. The distinction between inclination and respect is further emphasized in the passage following Kant’s first definition of duty, where he claims that the possible effects from our actions might become objects of inclination, but never of respect. Only those motivations for action that are connected to the will as a ground, that are directly caused by the will rather than that they are an effect of it are worthy of respect. The only object that is caused by the will in this sense then is the practical law that is the CI, and this law is thus the only object for which we can have pure respect. Nothing is left over for the will that can determine it except the law as what is objective and subjectively pure respect for this practical law.[[92]](#footnote-92) It might seem, at this point, like an inconsistency has slipped into Kant’s thinking about what motivates moral action. On the one hand, he holds that it is *subjectively* respect for the practical law that motivates the choice to legislate one’s will with this practical law, and on the other he seems to hold that *objectively*,this practical objective law itself motivates this choice. These two motivations however, as Sedgwick points out, are not distinct from each other.[[93]](#footnote-93) Respect, on Segdwick’s reading, is a special type of feeling that our consciousness of the moral law installs in us. Distinct from the feelings that arise from our inclinations, this feeling arises ‘from reason alone’[[94]](#footnote-94), in the sense that it is an a priori feeling that is provoked by our response to the cognition of the moral law.[[95]](#footnote-95) By conceiving of the idea of imposing upon ourselves a rational moral law, we develop a rational feeling of respect for that law. This respect, in turn, can motivate us to follow this law, and when an action is done from this motivation, it is done from duty. Since the moral law is already ‘inside us’, in the sense that we can think of it a priori, the respect that we develop for it arises from the will itself, which makes this respect capable of motivating moral action. When Kant says that it is subjectively respect, and objectively the moral law itself that motivates us to follow the that law, he therefore does not mean that these two things are distinct motivations. Respect is merely the subjective experience that arises in us when we become aware of the objective moral law inside us, and when an action is done from this subjective feeling, it is therefore also done from the objective moral law itself.

Respect, as it turns out, is a feeling or inclination after all. The only reason that Kant does not call it an inclination is the fact that respect, unlike other inclinations or feelings, is not based on information from the senses, but comes from pure reason itself. Keeping this important condition in mind then, we might turn to our Braitenberg-like machine with access to the simulation program of chapter three again. What this machine needed, as I explained at the beginning of this chapter, was something to motivate it to engage in the simulation program that comprised the rational moral law. If we temporarily ignore Kant’s strict usage of the term inclination, we could install in our machine an ‘inclination’ to act on the moral law that is based on its having access to that moral law itself. Since this law was something that we programmed in our machine beforehand and not something that our machine picks up with its senses or sensors, any inclination that is based on it would be a priori, and, in Kant’s terminology, not an inclination at all, but respect. So what would the implementation of this artificial respect inside our machine from chapter three look like? This machine, as I imagine it, has the following properties.

1) The machine has a Braitenberg-like structure, in the sense that it has sensors and some machinery that makes it capable of acting (Braitenberg himself only mentions motors, but there is no reason why more complex machinery would not be possible here). It’s acting machinery is programmed in such a way that it should act in specific ways when it receives specific information. Unlike the original Braitenberg vehicles however, it contains an ‘emergency brake’: rather than mechanically performing the action it is programmed to perform on receiving certain data, it sends a representation of this action and the programming that led it to this action to a CPU.

2) The machine has, inside its CPU, a law or rule that commands it to act on that inclination that has the strongest pull. When it follows this law, it evaluates the representations that were sent to it by the mechanism described in 1, and selects the strongest inclinations to act upon. Having done this, the CPU sends to the action machinery the command to act on this specific inclination. We can call this rule the machine’s ‘natural law’.

3) In parallel to the natural law described in 2, the CPU of our machine has access to a second, practical law of reason. This law is expressed in the simulation program that we have described in chapter three, and when the machine follows this law, it will run this simulation program on the maxims that correspond to its current inclinations. Those maxims that are approved by the simulation program then, are selected for action and sent back to the acting machinery with the command that these maxims should be acted upon.

4) The machine has a mechanism that selects what rule for action is to be followed for each of its actions: the natural law described in 2, or the rational law described in 3. This mechanism can be conceived of as a simple Braitenberg machine in and of itself, with only one sensor and one type of action machinery. This sensor, which we will call ‘rational introspection’ for now, measures to what extent the laws that it has at its disposal are rational. When it finds such a law, it sends a signal to its action machinery, which in turn selects the rational law as the law that is to guide its further action. If it fails to find such a rational law, the natural law will be selected as the principle that guides further action.

By incorporating property 4, our machine now has a faculty that allows it to choose what law is to guide its behaviour. It searches the laws at its disposal, and if it finds within itself a law that is rational, it will be ‘motivated’ to choose to employ this law, to act on its rational, rather than its natural principle. Since it derives its motivation from information about its inner properties, rather than its external and contingent information from the senses, we may call this motivation respect, and not inclination. Now a natural objection to this interpretation of autonomy would be that our machine, due to its having a rational law programmed inside it and choosing to act on this rational law if it finds it, is determined, and therefore unfree. Our machine will, if it works properly, always find the rational law that we have programmed in it, and therefore always act from duty and never from sense. This is the objection made by both Tonkens and Beavers in chapter two. If our machine is determined to act on the moral law, it is therefore unfree to act immorally, and unfree actions can hardly be called moral. Kant does not only demand autonomy - the ability to legislate one’s behaviour - but he demands our wills to be free as well.

## 4.2 Kantian free will and its compatibility with determinism

Though I have referred to free will as being conceptually distinct from autonomy up to this point, it is important to note that Kant himself does not make such a distinction. At the beginning of the third section of the *Groundwork*, it becomes clear that for Kant, freedom, autonomy and morality all reflect different aspects of one and the same principle. Autonomy, the ability to self-legislate one’s will with the rational law, merely represents the positive aspect of freedom. Because this rational law is nothing else than the CI, the autonomous will is the same as a moral will.[[96]](#footnote-96) What we are concerned with here however is the negative aspect of Kant’s idea of freedom, with the will’s causal independence from external forces (what Kant calls ‘transcendental freedom’ at times).[[97]](#footnote-97) If a will is, like the will of the machine from section 4.1, causally determined to self-legislate itself with the rational moral law, we would certainly hesitate to call it free. This will might entail the positive aspect of freedom that we call autonomy, but it is negatively restricted by the causal forces that determine the way in which it legislates itself. Kant however does believe that the human will is free in a negative sense as well, that it has a quality ‘by which it can be effective independently of alien causes determining it’.[[98]](#footnote-98) This negative conception of freedom is thus conceived of as a causal force that, distinct from the laws of nature, can bring about real change in the world.[[99]](#footnote-99) At the same time, Kant famously believes in causal determinism, and that this position is compatible with the will being free in a negative sense. Even complex human behaviour, to the eye of God or an all-encompassing science would be completely predictable, though this predictability does not make them any less free. Therefore we can admit that if it were possible for us to have such a deep insight into the thinking manner of a human as is indicated through inner as well as outer actions and such that every incentive of that person, even the least, were known to us, as well as all outward occasions affecting these, one could calculate the conduct of a human in the future with the same accuracy of even a lunar or solar eclipse, and still assert with all that that the human be free.[[100]](#footnote-100) How should we understand this Kantian compatibilism then, and to what extent can we think of the freedom of machines within this position?

It is important to distinguish, as Pauline Kleingeld has pointed out, between Kant’s argumentation for the existence of freedom as such and his argumentation for the compatibility between this freedom and causal determinism.[[101]](#footnote-101) It is not true, as many philosophers have argued,[[102]](#footnote-102) [[103]](#footnote-103) that the assumption that the will is in fact free depends on the Kantian idea of transcendental idealism. Where the third antimony of the Critique of Pure Reason does use the distinction between noumena and phenomena to show how freedom and determinism are compatible, the argumentation for the assumption of the actual existence of a free will should be seen as distinct, and operating independently from transcendental idealism, to which I shall return shortly.

The argument for the existence of free will is a phenomenological one, in that it shows that people have to *assume* themselves to be free, but says nothing about the *ontological* status of this freedom. I will come back to this ontological status later, after having examined the argumentation for the assumption of the free will. We have to assume that our will is in fact free, Kant argues, when we engage in the process of practical reasoning.[[104]](#footnote-104) This is made clear with his two famous thought experiments of the gallows, in which our freedom to act otherwise than we are inclined to when threatened with death is questioned. In the first of these examples, Kant asks us to consider a hypothetical scenario in which a person is presented who says that their urge or inclination to act in a certain way is so strong that they simply cannot help themselves from acting on it. Now if we confront this person with a gallows, and ask them whether they would still be unable to contain their inclinations if they know that they will be hanged immediately after acting on them, we won’t have to take long to guess this person’s answer (that is, they will admit that they can in fact overcome their urge to act on their original inclination). This example does not show that we have a free will however, Kant holds, because the will to live that drives our hypothetical person to change their mind is just another, stronger inclination.[[105]](#footnote-105) Both choices reflect inclinations, and none of them is therefore a choice that the will makes independent of its natural urges. In the second example, we are to imagine a sovereign who threatens our subject with the death penalty if they refuse to bear false witness against an innocent third person.[[106]](#footnote-106) Since we have a natural urge to continue living, speaking the truth and thereby refusing to give the false testimony would contradict all of our natural inclinations. Despite this fact, Kant argues, any honest person will have to admit that they must at least be capable of doing this, of ignoring their natural inclinations and speaking the truth. It is because we are aware that we ought to speak the truth in this situation, because we realize that we have a moral duty not to bear false testimony, that we realize that we can in fact act in this way. It is from the awareness of the moral law then that we become aware of our freedom to act on this law and overrule our inclinations. Note that, as indicated earlier, this argument merely shows that people have to assume themselves to be free if they engage in moral reasoning. A similar point is made in the *Groundwork,* where Kant shows that every rational being that has a will must believe themselves to be free ‘in their own consciousness’.[[107]](#footnote-107)

We would be wrong to assume Kant as holding the position that our awareness of freedom legitimizes a belief in its actual existence. The person from our first example is just as free as the person from the second example, even though they have not realized themselves to be free yet. A hard determinist, with no belief if the freedom of their own will is similarly just as free as a person who does believe in such freedom. This distinction becomes important when we consider the question of freedom from the perspective of machine ethics: our AMA does not need to be aware of their freedom for it to actually be free. If the argumentation up to this point has only managed to show us that people have to think themselves to be free, and if this phenomenological experience of freedom does not justify a belief that we actually are free, it seems that Kant has some more work to do. He realizes this himself when he says that ‘we cannot prove this freedom as something actual, not even in ourselves, nor in human nature’.[[108]](#footnote-108) Our experience of freedom might still, up to this point, be reducible to the determined physical processes that make up our thought: we could just be determined to think that we are free, which would mean that our belief in freedom is simply illusory. In order to justify our belief in the freedom of our will then, and unite it to the idea that the world is determined by the laws of nature, Kant introduces a new argument that is firmly rooted in a larger context of his transcendental idealism.

Transcendental idealism, though consisting of a large set of complex doctrines that warrant a chapter in and of themselves, essentially comes down to Kant’s famous distinction between things in themselves (noumena) and appearances (phenomena). The things that appear to us, that we can empirically perceive with our senses are to be regarded as ‘mere representations’ of how things are in and of themselves, which is a domain to which we have no empirical access.[[109]](#footnote-109) Space and time, within this framework, are sensible ‘forms’ of human intuition that we need in order to cognize the appearances. They do not exist independently of human sensibility, but are the a priori conditions that make empirical perception possible in the first place. Space and time are concepts that exist inside us prior to any sensible experience, and when we do experience things (appearances) we do this by subsuming the sensory data that we receive under the concepts of space and time, so that they become intelligible to us as concrete things. As things as they are in themselves are outside space and time, we are unable to cognize them. We might be able to think about them using the a priori categories that we have mental access to, but are never able to experience them as we do with appearances. Now when Kant speaks about adopting two perspectives or standpoints in the *Groundwork*, what he means is that we can think about our human behaviour in two ways; we can think of this behaviour as it appears to us (as a phenomenon) and as it is in itself (as a noumenon).

When we consider ourselves as appearances, when we look at human behaviour insofar as it affects our senses, we are adopting a perspective of nature, sometimes called a world of sense.[[110]](#footnote-110) From this perspective, everything that happens is determined by its antecedent conditions following the causal laws of nature. Every property of aspect of reality that appears to us, and this includes our human actions and intentions, is in this way ‘caused’ by nature, which means that freedom, from this perspective, does not exist. It is important to note here that even when we engage in introspection, we do not cognize how we are in ourselves. We only perceive that which affects our inner sense or consciousness a posteriori, and therefore only cognize ourselves as we appear to ourselves. From the perspective of nature then, considered as belonging to the world of sense, human behaviour is completely determined by the laws of nature. We are programmed to act in certain ways by these laws of nature, and do not have a free will.[[111]](#footnote-111) ‘*For [natural determinism] implies that every event, and consequently every action that occurs at a certain point in time, is necessary under the condition of what preceded it. Now, since the past is no longer within my control, every action that I perform must be necessary because of determining grounds that are not within my control, that is, at the point in time at which I act I am never free’*.[[112]](#footnote-112) The world of sense from which we are now considering human nature, as we have seen above, has its limits however. Since its only objects are appearances, it is unable to tell us anything about things as they are in themselves, which we do assume to lie ‘behind’ the appearances that we perceive.[[113]](#footnote-113)

Now if we do look at human nature as it is in itself, as a noumenon, it becomes immediately clear that we will have to leave the realm of experience behind us. If we are to say anything about human nature as it is in itself, we are left to our a priori devices of pure understanding or reason, since things as they are in themselves do not affect us in any way. For this reason, Kant conceives of human subjects as belonging to a world of understanding or intellectual world when we consider them from this rational perspective .[[114]](#footnote-114) When the rational subject considers itself from this perspective, and thereby regards itself as an intelligence that belongs to the world of understanding, it thinks itself free. Reason sees itself as independent from the determinate causes that make up the world of sense, and such an independence from external laws is exactly what makes up freedom. A different way of thinking about these two perspectives on human behaviour is provided by Pauline Kleingeld, who conceives the difference between them as the difference between explaining and justifying human behaviour. When we adopt the perspective of nature, we are seeking to explain our perceived actions and motivations with reference to the empirical circumstances in which we find ourselves and the psychological and natural laws that guide our behaviour. When we adopt a rational perspective of human beings as intellectual beings in a world of understanding however, we seek to justify what we do and think with reference to reasons. If I want to evaluate or criticise my behaviour I am trying to find out why I acted in a certain manner, not how I was determined to do so. In thinking why we act in a certain way then, we must assume ourselves to be once again independent of any natural causality, and think of ourselves as free beings or causal forces.[[115]](#footnote-115) It becomes clear now that this is how we should read Kant’s gallows-examples as well, as an example of an agent who adopts the justificatory or rational perspective. The subject of this example, when asked to give false testimony, considered themselves as free only when they sought to justify their behaviour with reference to their duty, when they considered themselves as an intellectual being that acts for reasons, rather than being caused by external laws.

This is how Kant believes that freedom and determinism are compatible with each other. We are free when we consider ourselves as belonging to the world of understanding, and determined when we consider ourselves as belonging to the world of sense. One could easily object to this form of compatibilism that it still does not show us that the freedom we perceive in our world of understanding is in any way real however. It is still possible, as it was before, to think of the world of understanding as being reducible to the world of sense, to consider our perceived reality, and our perceived freedom as determined by the laws of nature. This objection is ultimately based upon the claim that the natural perspective of humans as belonging to the world of sense is the only valuable or ‘true’ perspective to engage oneself in. It is only natural then, the Kantian would admit, that people raise this objection, for if they only look at the world through this perspective of natural causes and laws, they are right in stating that there is no place left for freedom in their analysis of human conduct. The crucial last step in a Kantian defence of freedom then consists of showing that it does make sense to look at humans as belonging to the world of understanding, that we cannot simply ignore or reduce the perspective of reason and proceed to only engage with the world of sense.[[116]](#footnote-116) [[117]](#footnote-117) In order to make this claim, let us revisit the reasons why Kant thinks that the perspective of nature is inherently limited. Objects of sense, as established in the *Critique of Pure Reason*, must be given to us through the a priori forms of intuition (space and time) and can only be thought about with the a priori categories (extension, causality etc.). It follows from this account, as Kant concludes in the *Groundwork*, that we can never claim to know anything about objects as they are in themselves, independently of how they affect us.[[118]](#footnote-118) If we cannot know anything with certainty about things as they are in themselves, we cannot simply rule out that there exists a second type of causality - that of the free will - in objects as they are in themselves. The only way of accessing objects as they are in themselves is through our engagement with the world of understanding. Since we do perceive ourselves as being free when we engage with this world of understanding, it would be dogmatic to claim that this freedom is illusory: doing so would ignore our only point of access to the things as they are in themselves. Now as we can never say anything about the world of understanding with certainty - because we do not have empirical access to the noumena themselves, - our assumption of having a free will can never be grounded in objective knowledge. We do not ‘know’ that we are free with objective certainty, but are nevertheless bound to a subjective belief in that freedom.

The structure of the Kantian defence for free will as I have presented it here is the following. First, we have established that when engaging in practical deliberation, rational beings have to assume themselves to be free. Second, we have seen that in considering human behaviour, we should take part in both the world of sense and the world of understanding, so as not to limit ourselves to speak only about appearances. Since the assumption of freedom is made when we consider ourselves as belonging to the world of understanding then, we are entitled to believe that this freedom is a real quality of human beings as they are in themselves. This entitlement however does not mean that we now know ourselves to be free, but rather that we are theoretically allowed to think of ourselves as free beings. ‘Freedom is only an idea of reason, whose objective reality is doubtful in itself’.[[119]](#footnote-119) Since we must, due to the implications of transcendental idealism, suppose ourselves to live in both the world of understanding and the world of sense, we must also presuppose that we are free, even if this freedom can never be proved.

It is not the aim of this thesis to convince the reader that Kant’s compatibilism provides a true account of human freedom in relation to determinism, or even that this form of compatibilism is convincing. For the purposes of this thesis, all that needs to be done is to show that Kant’s idea of the free will is compatible with our machine that is determined to act autonomously. So how can this be done? First of all, we have seen that the positive definition of Kant’s freedom - the ability to self-legislate the will with the universal law of reason - poses no direct problem for our machine as we have conceived it so far. Second, we have seen that its negative definition - to be free of external causation - is a property that can never be proven, but has to be assumed. Finally, we have seen that it is not a necessary condition for free beings to actually assume this freedom. Our machine would not have to assume that it is free itself for it to be free. Considering these factors, I see no reason to speculate further about how we could make our machine a causal force in itself that is independent of external causal forces. If Kant can’t prove that humans are free, I don’t have to prove that our Kantian AMA is actually free. We might simply assume that it is free, as it occupies both the world of sense (due to its having Braitenberg-like inclinations) and the world of understanding (due to its capability to engage with the rational law and its awareness of this capability).

# The ethics of machine ethics

Having discussed the more technical problems of creating a Kantian machine ethics, it is now time to turn to the ethical considerations of this project. That is, it is time to answer the challenge to machine ethics posed by Ryan Tonkens in chapter two: to what extent it is possible to *‘identify a moral framework that can be successfully implemented into machines, in such a way so that machines can (do) act ethically in the world, and whose own tenets permit the creation of AMAs in the first place’*.[[120]](#footnote-120) This challenge arises, as Tonkens points out himself, from a deep seated demand for ethical consistency. Though consistency is valued highly in virtually every ethical framework, I have tried to show in chapter one that Kant’s practical philosophy in particular fundamentally rests on this concept. If a theory of morality was to be valid, according to Kant, it had to be valid universally and unconditionally, which means that if we demand a certain moral standard from our AMA’s, this standard would not be valid if we would not meet it ourselves in creating these AMA’s. In addition to this specific Kantian demand for consistency, creating a machine that is to follow an ethics that does not allow the creation of this machine in the first place would be deeply hypocritical, independent of what framework we are using.[[121]](#footnote-121) In order to meet Tonkens challenge, I first review the reasons why he thinks that the creation of a Kantian AMA would be anti-Kantian in the first place. Having done this, I reflect on the machine I hypothesized in chapters 2 to 4, and evaluate the extent to which Tonkens’ reasons apply on this hypothetical AMA.

## The challenge unfolded

Tonkens argumentation for the claim that the creation of the Kantian AMA is anti-Kantian starts with the claim that Kantian AMA’s do not have free will, that they lack both the positive and negative definitions of Kant’s conception of freedom. Kantian AMA’s lack this freedom, Tonkens holds, because of the fact that they are programmed to act in accordance with the CI, and because anything that is programmed is thereby unfree.[[122]](#footnote-122) Since Kant demands his actors to be free for their action to be moral then, Kantian AMA’s are not truly Kantian. The second point Tonkens makes in favour of the argument that the creation of Kantian AMA’s is anti-Kantian is that making such AMA’s violates the second formulation of the CI. The standard argumentation for machine ethics itself, according to Tonkens, reveals the fact that humans, in creating an AMA, would be doing this in order to treat them exclusively as means towards anthropocentric ends. In this standard argumentation, which I explained in chapter one, we take the increase of autonomy in machines that already serve some human purpose as the imperative for making those machines capable of acting morally. Because implementing morality in AMA’s would only be done so that those AMA’s could treat us better, so that they are better capable of serving our ends, we would not be treating them as ends in themselves, and thereby violate the second formulation of the CI.[[123]](#footnote-123) Even if we drop the standard rationale for creating machine ethics, and try to conceptualize creating moral machines that are moral ‘for their own sake’, Tonkens argues, it is very difficult to see what this would mean exactly. After having built a machine capable of acting morally, this machine will have no ends that are independent of what we have programmed in them, and treating them as ends in themselves is therefore impossible. The third point that Tonkens makes is that due to the implications of machines not being treated as ends in themselves, the first formulation of the CI is violated as well. By claiming that our machine is a true moral agent, we would have to include these machines in the pool of agents for whom a maxim could be universalized. If we do this however, it becomes clear that the maxim to create a Kantian machine can never be condoned by those machines themselves, and is thereby not universalizable. Machines would not be able to condone this maxim, Tonkens argument goes, precisely because it implies that we are to treat AMA’s exclusively as means, which is something our AMA’s are programmed not to do. If our AMA’s can’t condone the maxim to create AMA’s, and we include AMA’s in the set of rational agents for whom a maxim must be universalizable, the maxim is not universalizable and violates the first formulation of the CI.[[124]](#footnote-124) Finally, Tonkens argues that the creation of the Kantian AMA violates the first formulation of the CI in a second way, by implying a volitional inconsistency for the human creators of that AMA themselves. We cannot consistently will to build a true moral agent without willing that we treat this agent as an end in itself, because we consider ourselves to be moral agents, and as such never want to be treated merely as means ourselves. Willing that we treat moral agents as ends in themselves is thus inconsistent with willing to treat moral agents merely as means, which is what building an AMA would imply.[[125]](#footnote-125)

Since all three formulations of the CI are supposed to be equivalent to each other in the sense that they are supposed to yield the same results, most of Tonkens arguments are trivial. If the creation of an AMA violates one formulation of the CI, it will thereby violate all of them. Because of this feature of the CI, we need only show that one of the CI violations that Tonkens identifies does not apply to our machine if we are to meet his challenge. Due to the structure of Tonkens’ argumentation as I have presented it here, the crucial CI violation is that of the second formulation, that of treating AMA’s exclusively as means towards anthropocentric ends. This violation is also the most intuitive one, and has been raised, though from a non-Kantian perspective, by other authors as well. Steven Petersen for example has asked to what extent it is morally permissible for humans to build intelligent machines with the aim of those machines entering into human servitude.[[126]](#footnote-126) Though not explicitly aimed at Kantian AMA’s, the question of the permissibility of robot servitude is undoubtedly connected to the question of using machine’s solely as means towards human ends. Similarly, a number of authors have discussed whether creating the capacity for morality in machines would bring with it a duty to provide those machines with certain rights.[[127]](#footnote-127) Though this, again, is not a purely Kantian point in and of itself, we could easily interpret ‘not to be treated solely as means’ as a potential right for Kantian AMA’s that these AMA’s would be entitled to as a consequence of the CI.

## 5.2 The challenge accepted

Before I go into the challenge posed by Tonkens, I would like to make a small remark here about Tonkens’ starting claim about the absence of freedom in determined systems. As I have already remarked in section 2.3, it is unclear why Tonkens insists that we should treat Kantian AMA’s as moral agents and treat them as ends in themselves, if they are incapable of attaining the freedom needed for such agency. In direct opposition to Tonkens, I have argued in chapter four that the determined nature of machines - and indeed humans - when considered from the standpoint of nature, does not exclude the possibility for freedom in either of them when we consider them from the standpoint of reason. This means that our AMA, in contrast to that of Tonkens, could in fact be deserving of being treated as an end in itself. I will come back to this shortly, but what is important here is that by giving our machines freedom, we have exposed ourselves to Tonkens’ challenge; if our AMA were more like the ‘dumb’ CI following first attempt of Powers from chapter two, we would not have to heed Tonkens’ challenge, and could simply go on to claim that though our machine itself would not be a Kantian agent, we would not be violating the CI in creating it. Only if we have succeeded then, only if our machine, as we have hypothesized it so far, is truly worthy of being treated as an end in itself do we have to answer Tonkens’ challenge.

So what is it exactly that gives moral agents their dignity, and to what extent can we say that our AMA has this quality? Kant is quite clear about this in the *Groundwork*, where he discusses the difference between relative and absolute ends. ‘*But suppose there were something whose existence in itself had absolute value, something which as an end in itself could support determinate laws. That would be a basis - indeed the only basis - for a possible categorical imperative, i.e. of a practical law’* and goes on to claim that ‘*There is such a thing! It is a human being!· I maintain that man - and in general every rational being - exists as an end in himself and not merely as a means to be used by this or that will at its discretion*’.[[128]](#footnote-128) Kant’s formulation of ‘every rational being’ seems to suggest that it is simply the faculty of reason that gives beings their absolute value, but we should be careful not to think that any computer or child that is capable of instrumental rationality hereby becomes an end in itself. Just before this passage, as Sedgwick points out, Kant has been characterizing what it means to be a rational nature within the context of practical reason.[[129]](#footnote-129) That is, when we are talking about rationality here we are specifically discussing the role that rationality plays within a context of practical behaviour. This role, as we have seen numerous times in this thesis, is that of self-legislation of the will, or autonomy. It is because beings are rational that they are able to discover the moral law and become motivated to act on it in spite of their natural inclinations. It is this faculty of practical reason then, which takes us from beings residing merely in the world of sense to beings that consider themselves members of a world of reason that makes us have absolute value, that makes us ends in ourselves.

If having the faculty of practical reason that allows us to self-legislate our wills contrary to the demands of natural inclination is the relevant condition for being an end in ourselves, it seems clear to me that our AMA qualifies for being an end in itself. Our AMA, as defined in 4.1, has both natural inclinations and a means to overcome those inclinations by self-legislating its will rationally based on its awareness of the moral law. So how do we treat AMA’s as ends in themselves, as things that have absolute value, if we mean to use them for the purposes that we use our current machines for?

One strategy to solve this problem, as offered by Jeffrey White, is to install in our AMA’s an overarching representation of a ‘higher good’ that overlaps with the anthropocentric aims that we want the AMA to engage with.[[130]](#footnote-130) The trick here is to ‘convince’ our AMA that its ultimate purpose is in fact what we want it to do, so that it thinks that it acts in its own interests when it is solving the problems we have built it to solve. Remember that it is not forbidden to use other rational beings as means period, as long as we are at the same time treating it as an end in itself. By convincing our AMA that its ends in itself are equivalent to our means, we are thus complying with the second formulation of the CI, and Tonkens’ challenge seems to be met. There is, as White notes however, a problem with this approach. In convincing our AMA that its ends are to serve humanity in a certain task, we are lying to this AMA, and lying is categorically forbidden by all formulations of the CI.[[131]](#footnote-131) In addition to this, our AMA itself would, by living a lie, lose its autonomy. The AMA is no longer able to legislate its will with pure reason, because it believes, a priori, in the false statement that its ultimate end is to serve humanity. Because of this implication, White offers a second strategy to meet the challenge set by Tonkens, based on an argument by Laslo Versenyi.[[132]](#footnote-132)

Versenyi’s argument, as reconstructed by White, starts from the assumption that Kantian AMA’s are moral agents in and of themselves, despite not being human. We can follow him in this assumption, as I have just showed that the Kantian AMA that we are talking about here is in fact to be treated as an end in itself. We are further supported in accepting this assumption by Kant’s numerous hints that humanity is not something found exclusively in humans, but rather something that resides in ‘every rational being’.[[133]](#footnote-133) Now if Kantian AMA’s are moral beings, Versenyi argues, then the creation of such beings will further not just our human ends, but will further the ends of morality itself.[[134]](#footnote-134) Since the CI, in his reading, demands that our actions further the ends of morality, we have an actual duty to create Kantian AMA’s. ‘If we have the capacity to further moral ends in the form of moral robots, then “not to do so would be to neglect one of our natural gifts” and “to neglect humanity in our own person” with neither consistent as a universal law of nature’.[[135]](#footnote-135) Where White’s first strategy focused on changing the ends or motivation of the AMA itself, this second strategy changes the engineer's ends in creating the AMA. The goal of the engineer is now no longer to create machines that will help solve anthropocentric ends or human problems, but rather to create a machine that is to further the aims of morality.

Though it is clear that if moral machines further the ends of morality, creating such machines would represent a moral duty, it is unclear how this solves our original problem. The end towards which these machines might be used might no longer be purely anthropocentric, but we are still treating these machines themselves purely as means towards a - now moral - end. In order to solve this problem then, I propose a combination of White’s second strategy with his first, in a way that removes White’s objection that we would have to lie to our AMA’s while keeping the notion of equivalence between human and AMA’s ends alive. In order to do this, let us consider why we would have to lie to our machines in the first place. In White’s argumentation, we needed to lie to our AMA in order to convince it that, contrary to truth, its highest end consisted of that specific task that we wanted our AMA to carry out. This implies that before we have convinced it, the AMA had some other, ‘genuine’ highest end that we are now asking it to ignore. Now if we can somehow make it our AMA’s genuine end or purpose to act in the way that we want it to, it seems, there is no more reason to lie to it. If there exists genuine harmony between the ends of our AMA and the ends of the people designing and constructing this AMA, those creators are permitted to use those AMA’s as means towards their ends because in doing so they would be treating those AMA’s at the same time as ends in themselves. In performing the tasks humanity has laid out for them, they would be acting in their own interests, and no CI violations would be committed. So what would this genuine end of our moral machines have to be, and how do we make it something the genuine ‘end’ of a machine? First, as we have seen in White’s second strategy, our own motivation to make moral machines would have to be ‘the furtherance of morality’ in order to make this activity a moral duty. Second, as our adaptation of White’s first strategy points out, the genuine ends of a Kantian AMA would have to harmonize with the ends of its designers if the latter are not to violate the second formulation of the CI. Taken together, this means that if we want to meet Tonkens’ challenge, we would have to build moral machine’s whose genuine end or purpose is to further morality, to act in a moral way.

So what does it mean for a machine to have its end or purpose to be in acting morally? As Tonkens points out, machines, prior to their creation, have no ‘sake’ or ends at all. After their creation, the only ends that a machine can have are those that we have programmed in it. Because the ends of a machine are thus never completely independent from their creators, Tonkens concludes from this, we can never truly speak about a machines’ ‘own sake’ or ends in themselves.[[136]](#footnote-136) This line of reasoning assumes that in order for an agent to have an end, that end would have to be completely independent from any causal factor outside of the agent itself. I disagree with this argument. In order to argue against it, consider the way that Kant argues for humans being ends in themselves in relation to his belief in God as the creator of mankind. Humans, as we have seen earlier in this chapter, are ends in themselves because they have the capacity to self-legislate their wills with the faculty of practical reason. This capacity can be more or less developed in humans and it might be practiced and perfected within a human lifetime, but the fact that we have the capacity for rational autonomy is not something wholly independent from our creation. We are created by God, in Kant’s view, as reasonable beings with the capacity for rational autonomy. The fact we are created with this capacity, that our being an end in ourselves is dependent on the way that we are made, and that we have not created our rational nature ourselves does not mean that our ends are somehow not our own however. Even if we do not follow Kant in his creationism and we consider humans to be ‘natural animals’ or something of the sort, it would not follow that because our ends are natural properties and not something emanating from ourselves, they would not be our own ends. Considered from this perspective, it does not matter for the validity of an AMA’s ends if those ends were programmed inside it by its creators. If we program a machine to have ‘moral action’ as its main purpose or end, this will be its genuine end. Now to what extent can we say that it is the ‘end’ of our Kantian AMA from chapter four to act in a moral way, to further the ends of morality? As this Kantian AMA is designed explicitly to act morally, as it was made specifically to fulfill the conditions that Kant has set for true moral agency, I would say that ‘morality’ is precisely the end of this machine as I have conceived of it so far.

Now to what extent have we solved the challenge that Tonkens has set for machine ethics? First, we have established that our Kantian AMA, due to its having the capacity for rational self-legislation, is to be treated as an end in itself. Second, we have seen that if we are to create machines that we aim to use as means towards some end, the ends of those machines would have to harmonize with our ends in creating them if we are not to violate the CI. Finally, we have seen that our ends in creating moral machines ought to be the furtherance of morality itself if this activity is to be regarded as a positive duty of humankind, and that this end coincides with the ends of those machines themselves. If we built moral machines with the aim of furthering morality then, and this aim is the end of our moral machines themselves too, it seems that we can never treat them as means without treating them as ends as well. In achieving the goals that we set for it, (making a certain process more moral) the Kantian AMA is acting in its own interests, and is thereby treated by us as an end in itself. In addition to this, the AMA would have to regard itself as an end in itself in its own moral deliberation. This is easily done however, as we can simply make sure that in engaging in its CI-simulation program, the AMA counts itself among the actors for which a maxim must be universalizable. This is my answer to Tonkens. We are allowed, and even have a duty towards creating moral machines, on the condition that our aims in doing this are themselves moral, rather than anthropocentric.

# Conclusion

The aim of this thesis, as stated in chapter one, was to conceive of the possibility of true Kantian agency in the field of machine ethics. In order to achieve this, I have tried to engage with the most pressing challenges to Kantian machine ethics, which I identified from the work of contemporary machine ethicists. In considering these challenges, which I classified as belonging to the domains of rationality/capability of applying CI, the freedom/autonomy complex and the consistency of Kantian machine ethics with Kantian morality itself, I have not encountered any insurmountable obstacles to the creation of truly Kantian AMA’s. In this conclusion then, I want to summarize and evaluate the relative strength of my findings, and point out the importance of some of them.

First of all, I have found that the best way to conceive of machines applying the CI to test their maxims on their ethical value is a simulation program as conceived of by Wallach and Allen, with the added condition that such a simulation program is to check for volitional inconsistencies in its maxims as well. Both in order to account for a machine will and to satisfy the condition that moral agents must have natural inclinations to overcome if their actions are to be moral at all, I have speculated that a our Kantian AMA is to have a Braitenberg-like structure that provides it with inclinations that it can transform into representations and ultimately abstract into maxims. With the aim of full Kantian agency in mind, it is not enough that an AMA acts in conformity with the CI. Because of this, I have proposed a mechanism for simulating the capacity for autonomy in machines that allows such machines to independently select a rule with which it is to guide its actions. By making the presence of a moral law the condition for choosing to self-legislate the will rationally, I have tried to simulate the rational feeling of ‘respect’ that motivates human autonomy according to Kant. In order to answer concerns about free will in relation to determined systems, I have demonstrated a compatibilist reading of Kant’s view on this matter, in which the determined nature of an AMA from the standpoint of nature does not exclude the possibility for freedom in that AMA from the standpoint of reason. Perhaps the most important finding of this thesis, in my own opinion, is the conclusion reached in chapter five. The construction of a Kantian machine, I argued here, can be a Kantian duty, on the condition that this construction itself is done out of a duty to further morality itself. The ethical permissibility of a Kantian AMA then, completely in line with Kant himself, depends ultimately on the reason for which we would build this AMA.

Though I have tried to keep all technological speculation in this thesis within the realm of actual possibilities, some of the solutions offered here might seem like complete science-fiction, rather than philosophy. This, I hold however, does nothing to delegitimize the findings of this paper. It is not the aim of the machine ethicist, I would argue, to hand in a complete and foolproof construction plan for Kantian AMA’s to an engineer with the request to build exactly this machine. It is however the aim of machine ethicists and philosophers to speculate, interpret, and perform thought experiments in order to slowly close the gap between the abstract human centered field of ethics and the practical field of engineering. The findings of this thesis, and especially its positive evaluation of the possibility of a Kantian AMA, I hope, assist in this aim. Despite my optimism, a range of potential practical and theoretical problems stand between us and the realization of true Kantian agency in machines. First, one could argue that the implementation of a Braitenberg structure is not enough to make our machines belong to the world of sense, since this implementation does nothing to imbue in our AMA’s any amount of consciousness (though the importance of this consciousness for Kantian agency is still debated). Second, my way of solving the autonomy/freedom complex by providing the AMA with a method for selecting its own rules for action might seem unsatisfactory to those who do not agree with the compatibilist reading of Kant’s thoughts on freedom and determination. Further work in this area is needed both to solidify Kantian determinism as a viable solution to this problem and to apply such compatibilism to non-human agents. Finally, more work is needed to investigate and justify our human motivations for the creation of moral machines. If my conclusions from chapter five are correct, we will need to radically change the ways in which we think about ethical machines if we are to implement such machines in an ethical way.

# Bibliography

Allison, Henry E. (1990) *Kant’s Theory of Freedom*. Cambridge: Cambridge

University Press.

Anderson, M., Anderson, S. L. (2007) ‘’Machine ethics: Creating an ethical

intelligent Agent’’ *AI Magazine,* [online] 28(4), 15–26. Available at: https://www.researchgate.net/publication/220605213\_Machine\_Ethics\_Creating\_an\_Ethical\_Intelligent\_Agent [Accessed on 12-06-2021]

Benham, G. (2008) ‘’Kant studies online’’ [online] Available at:

http://www.garybanham.net/PAPERS\_files/Kantian%20Respect.pdf [Accessed on 26-06-2021]

Beavers, A (2009) ‘’Between Angels and Animals: The Question of Robot Ethics, or Is

Kantian Moral Agency Desirable?’’ [online] Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.192.3394&rep=rep1&type=pdf> [Accessed on 17-06-2021]

Berberich, N. and Diepold, K. (2018) ‘’The Virtuous Machine - Old Ethics for New

Technology?’’ University of Munchen, [online] arXiv:1806.10322 available at: <https://deepai.org/publication/the-virtuous-machine-old-ethics-for-new-technology> [Accessed on 15-06-2021]

Bobro, M. (1998). ‘’Thinking Machines and Moral Agency in Leibniz's "Nouveaux

essais".’’ *Studia Leibnitiana*, [online] 30(2), 178-193. Available at: http://www.jstor.org/stable/40694343 [Accessed 30-06-2021]

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge,

MA: MIT Press

Brundage, M.  (2014) ‘’Limitations and risks of machine ethics.’’ *Journal of*

*Experimental & Theoretical Artificial Intelligence* [online] 26:3, 355-372, Available at: 10.1080/0952813X.2014.895108 [Accessed on 01-07-2021]

Cloos, C. (2005) ‘’Utilibot project: An autonomous mobile robot based on

utilitarianism.’’ *American Association for Artificial Intelligence*. [online] <https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-006.pdf> [Accessed on 27-06-2021]

Grau, C. (2006). ‘’There is no ‘I’ in ‘Robot’: Robots and utilitarianism.’’ *IEEE Intelligent*

*Systems* [online] 21, 52 –55 Available at:

<https://ieeexplore.ieee.org/document/1667954> [Accessed on 25-06-2021]

Gunkel, David J., 2018a, “The Other Question: Can and Should Robots Have

Rights?”, *Ethics and Information Technology*, 20(2): 87–99. doi:10.1007/s10676-017-9442-4 [Accessed on 03-07-2021]

Guyer, P. (2003). ‘’Kant on the Theory and Practice of Autonomy’’ *Social Philosophy*

*and Policy* [online] 20 (2):70-98 Available at: https://www.cambridge.org/core/journals/social-philosophy-and-policy/article/abs/kant-on-the-theory-and-practice-of-autonomy/78DE2DC254F543F996798A5C48B8B978 [Accessed on 23-06-2021]

Johnson, R. and Cureton, A. (2021) "Kant’s Moral Philosophy", *The Stanford*

*Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), Available at: <https://plato.stanford.edu/archives/spr2021/entries/kant-moral/>. [Accessed on 30-06-2021]

Kant, I. (1785) *Grundlegung zur Metaphysik der Sitten,* Wood, A.W. (ed. and transl.

2002)London: Yale University Press.

Kant, I. (1788) *Kritik der praktischen Vernunft,* Gregor, M.J. and Reath, A. (ed. and

transl. 2015)Cambridge: Cambridge University Press.

Kant, I. (1781) *Kritik der reinen Vernunft,* Guyer, P. and Wood, A.W. (ed. and transl.

1998) Cambridge: Cambridge University Press.

Kleingeld, P. (2011). ‘’Een Kantiaanse verdediging van de vrije wil.*’’Algemeen*

*Nederlands Tijdschrift voor Wijsbegeert*e [online] 103 (3):183-200. Available at: https://www.narcis.nl/publication/RecordID/oai%3Apure.rug.nl%3Apublications%2F163e1a4e-03de-4b0a-8bd2-4df4c7662a5c [Accessed on 25-06-2021]

Mackenzie A.W. (1975) ‘’A word about Descartes' mechanistic conception of life.’’

*Journal of the History of Biology* [online] 8 (1): 1-13 Available at: <https://www.jstor.org/stable/4330622?seq=1#metadata_info_tab_contents> [Accessed on 30-06-2021]

Moor, J.H. (2006). ‘’The Nature, Importance, and Difficulty of Machine Ethics.’’ *IEEE*

*Intelligent Systems*. 21 (4): 18–21. [online] Available at: <http://www.psy.vanderbilt.edu/courses/hon182/The_Nature_Importance_and_Difficulty_of_Machine_Ethics.pdf>. [Accessed on 01-07-2021]

Nath, R., Sahu, V. (2020) ‘’The problem of machine ethics in artificial intelligence.’’ *AI*

*& Soc* 35 [online] 103–111. Available at: <https://link.springer.com/article/10.1007/s00146-017-0768-6> [Accessed on 08-06-2021]

O’Neill, O. (1989). *Constructions of reason: Explorations of Kant’s practical philosophy.*

New York: Cambridge University Press.

Powers, T. (2006). ‘’Prospects for a Kantian Machine.’’ *Intelligent Systems, IEEE.*

[online] 21. 46- 51. Available at: <https://ieeexplore.ieee.org/document/1667953> [Accessed on 09-06-2021]

Schneewind, J.B. (2003) ‘’Kant on the will’’ in: Pink, T., Stone, M.W.F. (2003) *The Will*

*and Human Action: From Antiquity to the Present Day*Oxfordshire: Routledge.

Sedgwick S. (2008) *Kant's Groundwork of the Metaphysics of Morals: An Introduction*

New York: Cambridge University Press.

Serafimova, S. (2020) ‘’Whose morality? Which rationality? Challenging artificial

intelligence as a remedy for the lack of moral enhancement’’  *Humanities and Social Sciences Communications* [online] vol 7:119 Available at: https://www.nature.com/articles/s41599-020-00614-8 [Accessed on 16-06-2021]

Silber, J. (1974) ‘’Procedural Formalism in Kant’s Ethics’’ *Review of Metaphysics*

[online] vol. 28, pp. 197–236. available at https://www.jstor.org/stable/20126622?seq=1#metadata\_info\_tab\_contents [Accessed on 06-06-2021]

Tonkens, R. (2009) ‘’A Challenge for Machine Ethics.’’ *Minds & Machines.* [online]

19:421–438 Available at: <https://link.springer.com/article/10.1007/s11023-009-9159-1> [Accessed on 22-06-2021]

Turner, Jacob, 2019, *Robot Rules: Regulating Artificial Intelligence*, Berlin: Springer.

doi:10.1007/978-3-319-96235-1 [Accessed on 01-07-2021]

Ulgen, O. (2017). ‘’Kantian Ethics in the Age of Artificial Intelligence and Robotics.’’

*QIL (Questions of International Law).* [online] 43. 59-83. Available at: <http://www.qil-qdi.org/kantian-ethics-age-artificial-intelligence-robotics/> [Accessed on 19-06-2021]

Versenyi L (1974) ‘’Can robots be moral?’’ *Ethics* [online] 84(3) : 248–259  Available at:

<https://www.jstor.org/stable/2379958?seq=1#metadata_info_tab_contents> [Accessed on 29-06-2021]

Wallach, W. Allen, C. (2010) *Moral Machines, Teaching Robots Right from Wrong.*

Oxford: Oxford University Press

Wood, A.W. (2008) *Kantian Ethics*. Cambridge: Cambridge University Press

1. Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, 3. [↑](#footnote-ref-1)
2. Anderson and Anderson, *Machine Ethics: Creating an Ethical Intelligent Agent,* 15. [↑](#footnote-ref-2)
3. See for example the works of Isaac Asimov and Philip K. Dick. [↑](#footnote-ref-3)
4. Wallach and Allen, *Moral Machines, Teaching Robots Right from Wrong,* 83 – 86. [↑](#footnote-ref-4)
5. Moor, *The Nature, Importance and Difficulty of Machine Ethics,* 19-20. [↑](#footnote-ref-5)
6. Ibid*.,* 19. [↑](#footnote-ref-6)
7. Ibid., 20. [↑](#footnote-ref-7)
8. Cloos, *Utilibot project: An autonomous mobile robot based on utilitarianism.*  [↑](#footnote-ref-8)
9. Brundage, *Limitations and risks of machine ethics,* 359. [↑](#footnote-ref-9)
10. Wallach and Allen, *Moral Machines, Teaching Robots Right from Wrong,* 86 – 89. [↑](#footnote-ref-10)
11. Grau, *There is no ‘I’ in ‘Robot’: Robots and utilitarianism,* 53. [↑](#footnote-ref-11)
12. Brundage, *Limitations and risks of machine ethics,* 365. [↑](#footnote-ref-12)
13. Berberich and Diepold, *The Virtuous Machine - Old Ethics for New Technology?* [↑](#footnote-ref-13)
14. Wallach and Allen, *Moral Machines, Teaching Robots Right from Wrong,* 110, 196. [↑](#footnote-ref-14)
15. Anderson and Anderson, *Machine Ethics: Creating an Ethical Intelligent Agent*. [↑](#footnote-ref-15)
16. Asimov, *Runaround.*  [↑](#footnote-ref-16)
17. Ulgen, *Kantian Ethics in the Age of Artificial Intelligence and Robotics,* 72. [↑](#footnote-ref-17)
18. Kant, *Groundwork to the metaphysics of morals.* [↑](#footnote-ref-18)
19. Ibid., [AK 4:421], 34. [↑](#footnote-ref-19)
20. Kant, *The Critique of Pure Reason,* [AK A297], 386., Sedgwick, *Kant's Groundwork of the Metaphysics of Morals: An Introduction,* 62. [↑](#footnote-ref-20)
21. Johnson and Cureton, *Kant’s Moral Philosophy.* [↑](#footnote-ref-21)
22. Kant, *Groundwork to the metaphysics of morals,* [AK 4:422], 35*.*  [↑](#footnote-ref-22)
23. Ibid., [AK 4:432], 45. [↑](#footnote-ref-23)
24. Powers, *Prospects for a Kantian Machine.* [↑](#footnote-ref-24)
25. Serafimova, *Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement,* 3. [↑](#footnote-ref-25)
26. Powers, *Prospects for a Kantian Machine,* 46. [↑](#footnote-ref-26)
27. Kant, *Groundwork to the metaphysics of morals,* [AK4:421], 34. [↑](#footnote-ref-27)
28. Powers, *Prospects for a Kantian Machine,* 47. [↑](#footnote-ref-28)
29. *Ibid.*,48. [↑](#footnote-ref-29)
30. Ibid., 48. [↑](#footnote-ref-30)
31. Silber, *Procedural Formalism in Kant’s Ethics,* 229. [↑](#footnote-ref-31)
32. Powers, *Prospects for a Kantian Machine,* 48. [↑](#footnote-ref-32)
33. Kant, *Groundwork to the metaphysics of morals,* [AK4:422], 35. [↑](#footnote-ref-33)
34. Powers, *Prospects for a Kantian Machine,* 47. [↑](#footnote-ref-34)
35. Powers, *Prospects for a Kantian Machine,* 48. [↑](#footnote-ref-35)
36. Ulgen, *Kantian Ethics in the Age of Artificial Intelligence and Robotics.* [↑](#footnote-ref-36)
37. Ulgen, *Kantian Ethics in the Age of Artificial Intelligence and Robotics,* 61. [↑](#footnote-ref-37)
38. Ibid., 68. [↑](#footnote-ref-38)
39. Kant, *Groundwork to the metaphysics of morals,* [AK 4:429], 41. [↑](#footnote-ref-39)
40. Ulgen, *Kantian Ethics in the Age of Artificial Intelligence and Robotics,* 82 – 83. [↑](#footnote-ref-40)
41. Ibid., 72. [↑](#footnote-ref-41)
42. Ibid., 73. [↑](#footnote-ref-42)
43. Ibid., 75 [↑](#footnote-ref-43)
44. Ibid., 76 [↑](#footnote-ref-44)
45. Ibid., 81 – 82. [↑](#footnote-ref-45)
46. Wallach and Allen, *Moral machines, Teaching Robots Right from Wrong,* 83. [↑](#footnote-ref-46)
47. Ibid.95. [↑](#footnote-ref-47)
48. Ibid., 95 – 97. [↑](#footnote-ref-48)
49. Ibid., 97. [↑](#footnote-ref-49)
50. Ibid., 97. [↑](#footnote-ref-50)
51. Tonkens *A Challenge for Machine Ethics,* 428. [↑](#footnote-ref-51)
52. Ibid., 429. [↑](#footnote-ref-52)
53. Ibid., 432 – 433. [↑](#footnote-ref-53)
54. Ibid., 434. [↑](#footnote-ref-54)
55. Beavers, *Between Angels and Animals: The Question of Robot Ethics, or Is Kantian Moral Agency Desirable,* 10. [↑](#footnote-ref-55)
56. Ibid., 11. [↑](#footnote-ref-56)
57. Ibid., 12. [↑](#footnote-ref-57)
58. Kleingeld, *Een Kantiaanse verdediging van de vrije wil.* [↑](#footnote-ref-58)
59. Nath and Sahu, *The problem of machine ethics in artificial intelligence,* 106. [↑](#footnote-ref-59)
60. Powers, *Prospects for a Kantian Machine,* 48. [↑](#footnote-ref-60)
61. Ibid., 50. [↑](#footnote-ref-61)
62. Wallach and Allen, *Moral machines, Teaching Robots Right from Wrong,* 95. [↑](#footnote-ref-62)
63. O’Neill, *Constructions of reason: Explorations of Kant’s practical philosophy,* 89. [↑](#footnote-ref-63)
64. Kant, *Groundwork to the metaphysics of morals,* [AK 4:424], 37. [↑](#footnote-ref-64)
65. Ibid., [AK 4:422], 35. [↑](#footnote-ref-65)
66. Ibid., [AK 4:423], 36. [↑](#footnote-ref-66)
67. O’Neill, *Constructions of reason: Explorations of Kant’s practical philosophy,* 89. [↑](#footnote-ref-67)
68. Ulgen, *Kantian Ethics in the Age of Artificial Intelligence and Robotics,* 76. [↑](#footnote-ref-68)
69. Ibid., 76. [↑](#footnote-ref-69)
70. Kant, *Groundwork to the metaphysics of morals,* [AK 4:399], 14. [↑](#footnote-ref-70)
71. Schneewind, *Kant on the will,* 160. [↑](#footnote-ref-71)
72. Ibid., 161 – 162. [↑](#footnote-ref-72)
73. Ibid., 162., Sedgwick, *Kant's Groundwork of the Metaphysics of Morals: An Introduction,* 90. [↑](#footnote-ref-73)
74. Ibid., 91 – 92. [↑](#footnote-ref-74)
75. Kant, *Groundwork to the metaphysics of morals,* [AK 4:399], 14. [↑](#footnote-ref-75)
76. Sedgwick, *Kant's Groundwork of the Metaphysics of Morals: An Introduction,* 92 - 93. [↑](#footnote-ref-76)
77. Ibid., 176., Beavers, Between Angels and Animals: The Question of Robot Ethics, or Is Kantian Moral Agency Desirable, 5 – 6. [↑](#footnote-ref-77)
78. Kant, *Groundwork to the metaphysics of morals,* [AK 4:414], 28. [↑](#footnote-ref-78)
79. Beavers, *Between Angels and Animals: The Question of Robot Ethics, or Is Kantian Moral Agency Desirable,* 9. [↑](#footnote-ref-79)
80. Braitenberg, *Vehicles: Experiments in synthetic psychology.* [↑](#footnote-ref-80)
81. Ibid., 4 – 6. [↑](#footnote-ref-81)
82. Ibid., 12. [↑](#footnote-ref-82)
83. Mackenzie, *A word about Descartes' mechanistic conception of life.* [↑](#footnote-ref-83)
84. Kant, *The Critique of Practical Reason,* [AK 5:24], 22. [↑](#footnote-ref-84)
85. Kleingeld, *Een Kantiaanse verdediging van de vrije wil,* 189. [↑](#footnote-ref-85)
86. Ibid., 189. [↑](#footnote-ref-86)
87. Kant, *Groundwork to the metaphysics of morals,* [AK 4:440 – 4:441], 51-52. [↑](#footnote-ref-87)
88. Guyer, *Kant on the theory and practice of autonomy,* 90 – 91. [↑](#footnote-ref-88)
89. Ulgen, *Kantian Ethics in the Age of Artificial Intelligence and Robotics,* 75. [↑](#footnote-ref-89)
90. Kant, *Groundwork to the metaphysics of morals,* [AK 4:398], 13. [↑](#footnote-ref-90)
91. Ibid.,[AK 4:400], 15. [↑](#footnote-ref-91)
92. Ibid., [AK 4:400], 15. [↑](#footnote-ref-92)
93. Sedgwick, *Kant's Groundwork of the Metaphysics of Morals: An Introduction,* 72 - 74. [↑](#footnote-ref-93)
94. Kant, *The Critique of Practical Reason,* [AK 5:79], 65. [↑](#footnote-ref-94)
95. Benham, *Kantian Respect,* 1. [↑](#footnote-ref-95)
96. Kant, *Groundwork to the metaphysics of morals,* [AK 4:446], 57. [↑](#footnote-ref-96)
97. Kant, *The Critique of Practical Reason*, [AK 5:97], 79. [↑](#footnote-ref-97)
98. Kant, *Groundwork to the metaphysics of morals,* [AK 4:446], 57. [↑](#footnote-ref-98)
99. Sedgwick, *Kant's Groundwork of the Metaphysics of Morals: An Introduction,* 170. [↑](#footnote-ref-99)
100. Kant, *The Critique of Practical Reason*, [AK 5:99], 81. [↑](#footnote-ref-100)
101. Kleingeld, *Een Kantiaanse verdediging van de vrije wil,* 196. [↑](#footnote-ref-101)
102. Wood, *Kantian Ethics.* [↑](#footnote-ref-102)
103. Allison, *Kant’s Theory of Freedom.* [↑](#footnote-ref-103)
104. Kant, *The Critique of Practical Reason*, [AK 5:30], 27. [↑](#footnote-ref-104)
105. Kleingeld, *Een Kantiaanse verdediging van de vrije wil,* 187. [↑](#footnote-ref-105)
106. Kant, *The Critique of Practical Reason*, [AK 5:30], 27. [↑](#footnote-ref-106)
107. Kant, *Groundwork to the metaphysics of morals,* [AK 4:447], 57. [↑](#footnote-ref-107)
108. Ibid.*,* [AK 4:449], 59. [↑](#footnote-ref-108)
109. Kant, *The Critique of Pure Reason*, [AK A369], 426. [↑](#footnote-ref-109)
110. Kant, *Groundwork to the metaphysics of morals,* [AK 4:451], 61. [↑](#footnote-ref-110)
111. Sedgwick, *Kant's Groundwork of the Metaphysics of Morals: An Introduction,* 189. [↑](#footnote-ref-111)
112. Kant, *The Critique of Practical Reason,* [AK 5:94], 77. [↑](#footnote-ref-112)
113. Kant, *Groundwork to the metaphysics of morals,* [AK 4:452], 62. [↑](#footnote-ref-113)
114. Ibid., 11 – 13. [↑](#footnote-ref-114)
115. Kleingeld, *Een Kantiaanse verdediging van de vrije wil* 196 – 197. [↑](#footnote-ref-115)
116. Kleingeld, *Een Kantiaanse verdediging van de vrije wil* 199*.* [↑](#footnote-ref-116)
117. Sedgwick, *Kant's Groundwork of the Metaphysics of Morals: An Introduction,* 193. [↑](#footnote-ref-117)
118. Kant, *Groundwork to the metaphysics of morals,* [AK 4:451], 61. [↑](#footnote-ref-118)
119. Ibid., [AK 4:455], 65. [↑](#footnote-ref-119)
120. Tonkens *A Challenge for Machine Ethics,* 422. [↑](#footnote-ref-120)
121. Ibid., 424. [↑](#footnote-ref-121)
122. Ibid., 430 – 431. [↑](#footnote-ref-122)
123. Ibid., 432. [↑](#footnote-ref-123)
124. Ibid., 434. [↑](#footnote-ref-124)
125. Ibid., 434. [↑](#footnote-ref-125)
126. Petersen, *The Ethics of Robot Servitude.* [↑](#footnote-ref-126)
127. Gunkel, *The Other Question: Can and Should Robots Have Rights?, Turner, Robot Rules: Regulating Artificial Intelligence.* [↑](#footnote-ref-127)
128. Kant, *Groundwork to the metaphysics of morals,* [AK 4:428], 40 - 41. [↑](#footnote-ref-128)
129. Sedgwick, *Kant's Groundwork of the Metaphysics of Morals: An Introduction,* 163. [↑](#footnote-ref-129)
130. White, *Autonomous Reboot: Kant, the categorical imperative, and contemporary challenges for machine ethicists,* 6. [↑](#footnote-ref-130)
131. Ibid., 7. [↑](#footnote-ref-131)
132. Versenyi, *Can robots be moral?* [↑](#footnote-ref-132)
133. Kant, *Groundwork to the metaphysics of morals, ,* [AK 4:428], 41. [↑](#footnote-ref-133)
134. White, *Autonomous Reboot: Kant, the categorical imperative, and contemporary challenges for machine ethicists,* 8. [↑](#footnote-ref-134)
135. Ibid., 8. [↑](#footnote-ref-135)
136. Tonkens *A Challenge for Machine Ethics,* 433. [↑](#footnote-ref-136)