Deontological Machine Ethics

Thomas M. Powers
Department of Philosophy
University of Delaware
24 Kent Way
Newark, Delaware 19716
tpowers@udel.edu

Abstract

Rule-based ethical theories like Kant's appear to be promising for machine ethics because of the computational structure of their judgments. On one formalist interpretation of Kant's categorical imperative, for instance, a machine could place prospective actions into the traditional deontic categories (forbidden, permissible, obligatory) by a simple consistency test on the maxim of action. We might enhance this test by adding a declarative set of subsidiary maxims and other "buttressing" rules. The ethical judgment is then an outcome of the consistency test. While this kind of test can generate results, it may be vacuous in the sense that it would do no more than forbid obviously contradictory maxims of action. It is also possible that the kind of inference in such a rule-based system may be nonmonotonic. I discuss these challenges to a rule-based machine ethics, starting from the framework of Kantian ethics.

Introduction

Let us consider three problems of ethical importance concerning computers.

> 1) What methods are used to ensure that computers behave reliably—that is, precisely according to the specifications of their programs?
>
> 2) What methods will deter humans from using (well-behaved) computers for evil ends?
>
> 3) How do humans produce computers that will *themselves* refrain from evil, and perhaps promote good?

Resolving each of these problems is pressing, in some moral sense, because harm can come from any of these three sources.

The last problem is, on my view, the question of machine ethics. Its resolution seeks a simulacrum of ethical deliberation *in the machine*. Since moral philosophers have long been divided as to what counts as genuine ethical deliberation in humans, it will be no slight to the machine if all it achieves is a simulacrum. It could be that a great many of us do no better.

A deontological ethical theory is a good candidate for machine ethics because it supplies particular duties or directives for action that come from rules, and rules are (for the most part) computationally tractable. Among well-known deontological theories, Kantian ethics, or some subset of Kant's actual views, are thought to be more readily formalizable as a procedure that generates duties. Kant called this procedure the categorical imperative. I will explore a version of deontological ethics along the lines of Kantian formalist ethics, both to suggest what computational structures would be required by such a view, and to see what challenges remain for a successful implementation of it. In reformulating Kant for the purposes of machine ethics, I will consider three views of how the categorical imperative works: mere consistency, common-sense practical reasoning, and coherency.

Kantian Formalist Ethics

In his *Grounding of the Metaphysics of Morals* (Kant 1981), Kant claims that the first formulation of the categorical imperative supplies the following ethical rule:

> Act only according to that maxim whereby you can at the same time will that it should become a universal law.

Kant goes on to explain that the moral agent is to consider each maxim (or plan of action) as though it were to belong to an entire system of maxims making up a prospective system of moral laws (or duties) for all rational beings. If we are to see the categorical imperative as a rule for a

machine that deliberates in an ethical sense, we must consider how it might offer a procedure for generating duties. The procedure of deriving duties from the rule—if we are to believe Kant—requires no special moral or intellectual intuition that might be peculiar to human beings.

> Therefore I need no far-reaching acuteness to discern what I have to do in order that my will may be morally good. Inexperienced in the course of the world and incapable of being prepared for all its contingencies, I only ask myself whether I can also will my maxim should become a universal law.

How then can agents know what to do? For a formalist Kantian, this decision problem is the same for humans as for machines. Kant himself, twenty years prior to the *Grounding*, sketched the following answer as another way to view the procedure of the categorical imperative (Kant 1991):

> If contradiction and contrast arise, the action is rejected; if harmony and concord arise, it is accepted. From this comes the ability to take moral positions as a heuristic means. For we are social beings by nature, and what we do not accept in others, we cannot sincerely accept in ourselves.

Preliminary Modifications

Since I do not here intend to offer a strict interpretation of Kant's ethics, I will set out only minimal scholarly requirements, and will focus mostly on the logic of a computable categorical imperative.

Recall that the first formulation of the categorical imperative demands that we act only on maxims which can be universally willed. Hence we must first determine which of our maxims are universalizable. One way to do this would be to build the theory of forbidden maxims $\boldsymbol{F}$, and then check to see whether any prospective maxim m $\boldsymbol{\epsilon F}$. The theory will be finitely axiomatizable if and only if it is identical to the set of consequences of a finite set of axioms. Though we suppose that ethics is complicated, the theory $\boldsymbol{F}$ should still be finite.

An affirmative answer to the question "is $m \boldsymbol{\epsilon F}$?" would be the outcome of the categorical imperative procedure on a forbidden maxim. But we need tests that will also generate the other two deontic categories of permissible maxims and obligatory maxims. Since there is only one other answer to the question above ("no") and two categories to fill out, the theory $\boldsymbol{F}$ alone would be a semidecidable—there would remain the possibility that the "no" answer might not be rendered. This follows from the completeness of first-order theories, which presumably $\boldsymbol{F}$ is.

Since maxims are "subjective principles of volition" or plans, the first formulation serves as a test for turning plans into instances of objective moral laws. This is the gist of Kant's notion of self-legislation; an agent's moral maxims are instances of universally-quantified propositions which *could* serve as moral laws--ones holding for any agent. Since we cannot specify at the outset the class of universal moral laws, we build the theory of machine ethics with individual maxims by applying the universalization step and then classifying the maxims into the deontic categories according to the results of the categorical imperative.

How should we understand the universality of the ethically approved maxims? A universalizable maxim is *not* one in which we can imagine everyone doing the "numerically" same act, i.e., repeating the very act in question. Likewise, a maxim does not fail because of the physical impossibility of reiteration or instantiation. Consider the following impermissible maxim and its universalized versions.

m: I will to kill John because I hate him.

U_1: Everyone who hates someone else kills him or her.
$(\forall x)(\exists y)(Hxy \rightarrow Kxy)$

U_2: Anyone who hates John kills him.
$(\forall x)(Hxj \rightarrow Kxj)$

Let us assume that the test should fail *m*, since killing out of hatred is immoral. U_1 is not an adequate universalization of *m* even though it would fail the universalization test, i.e., it could produce a contradiction. It would do so, however, on the trivial grounds that two people might hate the same person and hence one of them would be unable to complete the act. Of course the substitution of the killing with the mere beating of the object of hatred shows that it is not the reiteration aspect of U_1 that makes it morally objectionable; everyone who hates a particular person could beat that individual. U_2, on the other hand, could be universalizable even under the iteration condition. Someone might believe that only he hates John (perhaps because very few know him), but that his hatred is justified because anyone who knew John would hate him. So the "rigged" symbolization would let U_2 pass the test. On this analysis U_2 is acceptable only because of some particular facts about John. And of course no universalization of *m* should make it an acceptable maxim.

Accordingly, an additional condition on the logical forms of maxims must be that that the universalization test will quantify over circumstances and purposes as well as agents. If we do not have this restriction, it is possible that some maxims might be determinate with respect to either

the circumstance or purpose, i.e., some might be pure existentials like 'I intend to thank Dan for comments'. We do not want to interpret universalization in such a way that the procedure asks of that maxim "Can I will that everyone thank Dan for comments?" Hence the test should turn the maxim above into the question "Can I will that anyone thank *any other* person who, as a matter of fact, helps them with comments?"

Mere Consistency

The section above tells us what will count as a properly formulated input for the test: a maxim over which circumstances, purposes, and agents have been quantified—the last being universally quantified. A computer must have the ability to parse such categories from programmed ontologies, or else it will simply have to accept properly formulated input maxims. (Would the latter violate the condition that a machine ethics is deliberation by the machine itself?) To see whether the input is an instance of a moral law, and exactly what deontic category it belongs to, Kantian formalism assumes that the test of the categorical imperative is an algorithm that alone will determine the classes of obligatory, forbidden, and permissible actions. Let us suppose that, after the quantification step, the machine has universalized maxims of the following sorts.

1) $\forall z \exists x \exists y(Cx \,\&\, Py) \rightarrow Az$ *A is obligatory for z*
2) $\forall z \exists x \exists y(Cx \,\&\, Py) \rightarrow \neg Az$ *A is forbidden for z*

3) $\neg\, \forall z \exists x \exists y(Cx \,\&\, Py) \rightarrow Az$
and *A is permissible for z*
$\neg\, \forall z \exists x \exists y(Cx \,\&\, Py) \rightarrow \neg Az$

where Cx=x is a circumstance, Py=y is a purpose, and Az=z commits action A.

We now have definitions for the three deontic categories, though admittedly we have no account of superogatory action, or action "beyond the call of duty." Intuitively, we say that anyone in a particular circumstance with a particular purpose ought to A in case 1), anyone ought to refrain from A in case 2), and anyone may do or refrain from A in case 3).

A major defect in the simple account above comes when the machine goes beyond formulating the maxims in their universalized forms. It needs to be able to test the maxims for contradictions, but the only contradictions which can arise are trivial ones. This is so even when we take the theory of maxims to be closed under logical consequence. A robust version of the test, on the other hand, requires that the machine compare the maxim under consideration with *other* maxims, principles, axioms, intentions, etc. Obviously, the simple account of mere consistency will not do. It must be buttressed by adding other principles or axioms, in comparison with which the maxim can be tested for contradiction.

Commonsense Practical Reasoning

We can buttress Kant's mere consistency rule by adding a background theory Γ, against which the test can have non-trivial results. What does Γ look like? For Kantians, what goes into Γ usually depends on the line of interpretation one has for Kant's ethics generally. That is, the tradition is to supplement Kant's moral theory with principles from the two *Critiques*, the *Metaphysics*, or his shorter essays.

Kant's illustrations of the categorical imperative (Kant 1981) suggest that he adopts this very plan of buttressing the "mere consistency" version of the test. In these illustrations, Kant introduces some common-sense principles. For instance, he argues that, since feelings are purposeful, and the purpose of the feeling of self-love is self-preservation, it would be a contradiction to commit suicide out of the feeling of self-love. This may appear to us now to be a bad argument, but that is not to the point. Many contemporary Kantians have suggested that we ought to include in Γ some common-sense rules. Such rules are called, variously, postulates of rationality (Silber 1974), constraining principles of empirical practical reason (Rawls 1980), and principles of rational intending (O'Neill 1989). These are presumably non-trivial, non-normative axioms which somehow capture what it is to act with practical reason.

What we get when we build Γ with commonsense reasoning postulates is something that is probably closer to ethical deliberation in humans, but presents difficulties insofar as we do not have programs that implement commonsense practical reason. On the other hand, it is quite thin, morally speaking, to think that the consistency of a single universalized maxim is enough to recommend it. With the help of Γ, we can revise the test of the categorical imperative. A maxim is defined as "unreasonable" just in case it produces a contradiction, given the postulates of Γ. With the right postulates, the formal categorical imperative plus the maxim might yield good results. Of course, the definition and choice of postulates does no more than *stipulate* what counts as practical reason. The inclusion of any postulate in Γ is underdetermined by logical considerations alone.

Postulates of commonsense or practical reason do not share the logic of scientific laws or other universal generalizations. One counter-example is enough to disprove a deductive law, but commonsense postulates must survive the occasional defeat. The postulates of Γ, then, might be expressed in a nonmonotonic theory of practical reasoning.

Nonmonotonic logic is an attempt to formalize an aspect of intelligence, artificial or human. Nonmonotonic reasoning is quite commonplace. Consider that classical first-order logic is monotonic. If a sentence α can be inferred from a set of premises Δ, then it can also be inferred from any set Σ that contains Δ as a subset.

Nonmonotonic inference simply denies this condition because the larger set may contain a sentence that "defeats" the inference to α.

For example, the addition of 'Fritz is a cat' to a set already containing 'All cats are mammals' licenses the monotonic inference 'Fritz is a mammal'. But if we replace our deductive law about cats with a default rule, such as 'Cats are affectionate', we can see some conditions under which the inference to 'Fritz is affectionate' would be defeated—namely, by additional information to the effect that 'Fritz is a tiger'. At the least, all bets should be off as to whether Fritz is affectionate.

While there are different ways to formalize nonmonotonic reasoning, we want to choose a way that will build on the categorical imperative as a consistency check for maxims. Reiter's default logic (Reiter 1980) is a good candidate for these purposes. The default rule of the example above on his account becomes

(1) If Fritz is a cat, **and it is consistent that Fritz is affectionate**, then Fritz is affectionate

The bold clause can be defeated by any number of additional facts, such as 'Fritz had a bad day', 'Fritz had a bad kittenhood', 'Fritz is a person-eater', etc.

Reiter suggests the following symbolization for this default rule:

2) $\frac{C : A}{A}$

C is here the precondition of the default, A the justification (in this instance), and A the consequent. This is a *normal* default rule since the justification is the same as the conclusion we are allowed to draw. The key to Reiter's default logic is the notion of an extension. Intuitively, an extension of a theory Æ(T) is a set of consequences of the default theory T=<W, *D*>, where W is a set of sentences and D the set of default rules. An instance of a consequent of the rules can be used in a consistency test if we can prove the precondition from the set of sentences W, and if the justifications are consistent with all of the consequents of the default rules that are used (Poole 1994). The extension of the theory adds to it all those default conclusions consistent with W.

The definition of an extension maintains the requirement of nonmonotonicity. Given a set of first-order sentences, adding the conclusions of default rules--and only the ones consistent with the sentences already in the theory--yields no conclusions not already in the extension. Introducing contradictions is the result that default extensions avoid. Default rules yield to facts; the rules are defeated but not vanquished. In monotonic logic, the universal laws are vanquished by counter-examples.

Kant seems to recognize that defeasible reasoning plays *some* role in moral thinking, and in this respect he is far ahead of his time. Kant (1981) writes that

> Since we regard our actions at one time from the point of view of a will wholly conformable to reason and then from that of a will affected by inclination, there is actually no contradiction, but rather an opposition of inclination to the precept of reason (*antagonismus*). In this the universality of the principle (*universalitas*) is changed into mere generality (*generalitas*), whereby the practical principle of reason meets the maxim halfway.

When we look closely at Kant's illustrations, we see that there is textual evidence to support the inclusion of default rules in Γ. We will look at two of Kant's illustrations of the first formulation.

Against Suicide

Kant offers the following account of moral deliberation for the person contemplating suicide. What I take to be an indication of his nonmonotonic reasoning is placed in bold.

> His maxim is 'From self-love I make it my principle to shorten my life if its continuance threatens more evil than it promises pleasure'. The only further question to ask is whether this principle of self-love can become a universal law of nature. **It is then seen at once that a system of nature by whose law the very same feeling whose function is to stimulate the furtherance of life should actually destroy life would contradict itself.** (Bold is added)

The default rule concerns the function or purpose of self-love, premise 3. below. The reconstructed argument runs as follows.

1. Anyone in pain and motivated by self-love (circumstance) shall try to lessen pain (purpose) by self-destruction (action).
2. Feelings have functions.
3. Self-love serves the function of self-preservation.
4. Self-destruction is the negation of self-preservation.

Therefore

5. A maxim of suicide is contradictory and hence the action is forbidden.

Premise 1. is much clearer if we symbolize it informally, allowing for the substitution in 4. (sd = ¬sp) and ignoring quantifiers.

1.1 If motivated by self love (sl) and a desire to lessen pain (lp), then negate self-preservation (¬sp).

Premises 2 and 3 will have the following modifications:

2.1 If one has a feeling and that feeling serves a

rational function, then it is rational to follow that feeling.

3.1 If self-love, and self-preservation is consistent,then self-preservation.

or formally in default logic

3.2 $\frac{sl : sp}{sp}$

Here the default nature of Kant's reasoning is obvious. Self-preservation is no universal law for Kant; it can be defeated under the right circumstances. Defeating conditions might include voluntary submission to punishment, sacrifice for loved ones, or stronger duties under the categorical imperative. But lacking those defeating conditions, and provided that the agent satisfies the conditions of the antecedent, it would seem that the universalized maxim plus the default rules yields the following argument.

1.1 $sl \& lp \rightarrow \neg sp$
3.2 $\frac{sl : sp}{sp}$

Hence

sp and $\neg$sp.

So we have the contradiction that the universalization procedure is intended to elicit.

Unfortunately, we cannot be entirely satisfied with this result. We must be careful here not to violate Reiter's conditions on extensions of default theories. If '¬sp' is included in the extension of T, then the default inference in 3.2 (sp) could never be added to T, given that it would not be consistent with the antecedent fact (¬sp) which is the defeating condition of the rule. We cannot get a contradiction in this seemingly straightforward way. The key to the problem is that Reiter's default rules will not introduce inconsistency into a set of sentences. If the conditions (facts) for inconsistency already obtain in the original set, the defeating condition is disallowed.

Against False-Promising

Kant (1981) gives the following account of how the categorical imperative would work with an input maxim of false promising, or promising repayment of a loan without the intention to repay.

> For the universality of a law that every one believing himself to be in need can make any promise he pleases with the intention not to keep it **would make promising, and the very purpose of promising, itself impossible, since no one would believe he was being promised anything.** (Bold added)

The traditional criticism of this illustration is that promising and borrowing would *not* in fact be impossible if false promising became a universal practice in the closely-defined circumstance of need. All that would follow is extreme caution in lending and an insistence on collateral. While such a result would be unfortunate and deplorable, it is consistent with other practices.

I do not believe this objection holds, however, because it misses the defeasible nature of both promising and lending. The institution of promising is dependent on two default rules, one for the debtor and one for the creditor. They are that promises are believed and that promises are kept. Both rules are occasionally defeated, and the prevalence of defeat threatens the institution. The "commonsense" creditor will not believe a promise after the debtor defeats the rule repeatedly. Likewise, the "commonsense" debtor knows better than to offer a promise to a rightly-incredulous creditor. But this is not to say that any one defeat of the rule of sincere promising threatens the institution of promising as a whole. Both creditors and debtors survive violations of the rules and continue to uphold the institution. What is clear, though, is that the monotonic understanding of the rule of promising--the understanding which renders the rule as the universal generalization "All promises are kept (or promising is destroyed)"--does not properly interpret the institution. The actual institution of promising depends as much on *surviving* a defeating instance as it does on the prevalence of non-defeat. So a nonmonotonic interpretation of the illustration makes sense of the practice, while the monotonic interpretation does not.

Difficulties for the Nonmonotonic Approach

There is a serious problem for the nonmonotonic approach to deontological machine ethics. Nonmonotonic inference fails the requirement of semidecidability. So it is not even the case that the nonmonotonically-enhanced categorical imperative is guaranteed to give a "yes" answer to the question: 'Is the action of this maxim forbidden?'. It is also not guaranteed to give a "no" answer. The obvious question, we might think, is: What good is the nonmonotonic categorical imperative?

Let me summarize my general argument. The nonmonotonic account of Kant's examples gives a better interpretation of the procedure than anything offered by traditional formalists using monotonic logic. We seem to need a background theory of common sense rationality in order for the categorical imperative test to give non-trivial results. Commonsense reasoning is not entirely captured by monotonic logic; and Kant himself, when he does provide clues as to the "buttressing" principles he assumes, gives us principles which can only make sense if they are defeasible. But this revised interpretation still fails an important *formal* requirement for machine ethics:

semidecidability.

Coherency

We might find some hope in our search for machine ethics by recalling what Kant is looking for: a *coherent system* of maxims. It will help if we return to illustrations of the categorical imperative in the *Grounding.*

The latter two illustrations (of allowing one's talents to rust and giving to others in need) compare the maxims to be evaluated with other maxims that are assumed or implied in the context of testing. In order to get the illustrations to work, they must be read as follows. A maxim of allowing one's talents to rust conflicts with what every rational being wills (according to Kant): the development of one's talents. If you will help from others when you are in need, you must will to help others when they are in need. What these cases share is the prohibition against acting on a maxim which is incoherent, *given* a system of other maxims. The other maxims which by comparison provide the coherency constraint are not "transcendentally" imposed. They are one's *own* maxims. Presumably, a machine could build a database of its own maxims.

Let us consider the procedure of the categorical imperative as a kind of "bottom-up" construction. Acting morally should be like building a theory, where the sentences of the theory are one's own maxims plus any consequences of the maxims. Call this theory *G.* The sentences of *G* are all first-order indicative. The theory also has two rules: R-in and R-out. For any maxim m_i, in M, which is considered, R-in says that m_i is allowed in M iff m_i and *G* are consistent.

What about maxims which we have acted upon in the past that turned out not to be permissible? Handling such incoherencies is analogous to the problems of belief revision explored by Gärdenfors (1988). If we allow the "impermissible" maxims to remain in *G*, it will automatically be an inconsistent set of sentences and hence the coherency constraint breaks down. Surely Kant does not insist on previous moral perfection in order to achieve a good moral disposition.

We can stipulate that *G* contains a rule (R-out) for excluding sentences which allow M to have contradictions. There is nothing mysterious about R-out. On the assumption that some maxim m_i turned out to be morally wrong, m_i and *G* are inconsistent, and $m_i \notin$ M. R-out serves the role of a confession of sins for the moral disposition.

There is one interesting aspect of *G* which poses a difficulty for a Kantian machine. Consider the limiting case where m_1 is the only member of M. We might call this the case of the moral infant. This first maxim had to be allowed to enter *G* by R-in, since by hypothesis *G* is empty and so it is consistent with everything. Now suppose the moral infant wants to test a second maxim m_2, and m_1 and m_2 are inconsistent. R-in disallows m_2, the violating maxim, but we cannot explain why it and not m_1 is impermissible, except to appeal to temporal priority. This seems irrational.

The problem with the limiting case m_1 holds not only for the first maxim but also for the nth maxim to be added to *G,* m_n. What reason other than temporal priority can we give for keeping the whole set of prior maxims, and disallowing m_n? There are of course good practical grounds for holding to the set of maxims one has already accumulated. Moreover, we might think that no moral agents are moral infants because everyone has at any given time an established set of maxims. But these objections are not to the point. If we are to construe Kant's test as a way to "build" a set of maxims, we must establish rules of priority for accepting each additional maxim. We must have what Gärdenfors calls an *epistemic commitment function*, though ours will be specific to moral epistemology. This problem is a species of the more general problem with anti-foundationalist epistemology.

The problem of the moral infant shows that a Kantian formalism in the constructivist or "bottom-up" tradition cannot build a coherent moral theory from nothing. A deontological theory must give reasons why the machine should not throw out an entire collection of maxims to allow entry of one otherwise "incoherent" maxim, m_n . More concretely, a Kantian theory must tell the agent who has "compiled" a good moral disposition why he (or it) may not *now* defeat all of those prior maxims and turn to a life of vice. I think there are many good reasons a Kantian could give, but they are not the ones that can be given in a "bottom-up" constructivist theory.

References

Kant, I. 1981. *Grounding for the Metaphysics of Morals*. Trans. J. Ellington. Indianapolis: Hackett. (First published in 1785)

Kant, I. 1991. *Bemerkungen in den 'Beobachtungen über das Gefühl des Schönen und Erhabenen'*. Hamburg: Felix Meiner. (Unpublished notes by Kant written 1764-66)

O'Neill, O. 1989. *Constructions of Reason*. Cambridge: Cambridge University Press.

Poole, D. 1994. Default Logic. *In Handbook of Logic in Artificial Intelligence and Logic Programming*. Edited by Gabbay, D., Hogger, C., and Robinson, J. New York: Oxford.

Rawls, J. 1980. Kantian Constructivism in Moral Theory. *Journal of Philosophy*. 77:515-72.

Reiter, R. 1980. A Logic for Default Reasoning. *Artificial Intelligence*. 13:81-132.

Silber, J. 1974. Procedural Formalism in Kant's Ethics. *Review of Metaphysics*. 28:197-236.