Prospects for a Kantian Machine

Thomas M. Powers, *University of Delaware*

One way to view the puzzle of machine ethics is to consider how we might program computers that will *themselves* refrain from evil and perhaps promote good. Consider some steps along the way to that goal. Humans have many ways to be ethical or unethical by means of an artifact or tool; they can quell a senseless riot by broadcasting a speech on television or use a hammer to kill someone. We get closer to machine ethics when the tool is a computer that's programmed to effect good as a result of the programmer's intentions. But to be ethical in a deeper sense—to be ethical in themselves—machines must have something like practical reasoning that results in action that causes or avoids morally relevant harm or benefit. So, the central question of machine ethics asks whether the machine could exhibit a simulacrum of ethical deliberation. It will be no slight to the machine if all it achieves is a simulacrum. It could be that a great many humans do no better.

Rule-based ethical theories like Immanuel Kant's appear to be promising for machine ethics because they offer a computational structure for judgment.

Of course, philosophers have long disagreed about what constitutes proper ethical deliberation in humans. The *utilitarian tradition* holds that it's essentially arithmetic: we reach the right ethical conclusion by calculating the prospective utility for all individuals who will be affected by a set of possible actions and then choosing the action that promises to maximize total utility. But how we measure utility over disparate individuals and whether we can ever have enough information about future consequences are thorny problems for utilitarianism.

The *deontological tradition*, on the other hand, holds that some actions ought or ought not be performed, regardless of how they might affect others. Deontology emphasizes complex reasoning about actions and their logical (as opposed to empirical) implications. It focuses on rules for action—how we know which rules to adopt, how we might build systems of rules, and how we know whether a prospective action falls under a rule. The most famous deontologist, Immanuel Kant (1724–1804), held that a procedure exists for generating the rules of action—namely, the *categorical imperative*—and that one version of the categorical imperative works in a purely formal manner.

Human practical reasoning primarily concerns the transformation between the consideration of facts and the ensuing action. To some extent, the transformation resembles a machine's state changes when it goes from a set of declarative units in a database to an output. There are other similarities, of course—humans can learn new facts that inform their reasoning about action, just as machines can incorporate feedback systems that influence their outputs. But human practical reasoning includes an intervening stage that machines (so far) seem to lack: the formation of normative claims about what is permissible, what one ought to do, what one is morally required to do, and the like. It's plausible that normative claims either are ethical rules themselves or entail such rules. These normative claims aren't independent of facts, and they don't necessarily lead humans to action. In fact, humans suffer from "weaknesses of the will," as Aristotle called them, that shouldn't be a problem for a machine: once it reaches a conclusion about what it ought or ought not to do, the output will follow automatically. But how will the machine reach the middle stage—the normative conclusions that connect facts to action through rules? I think this is the problem for machine practical reasoning.

A rule-based ethical theory is a good candidate for the practical reasoning of machine ethics because it generates duties or rules for action, and rules are (for

1541-1672/06/$20.00 © 2006 IEEE
Published by the IEEE Computer Society

the most part) computationally tractable. Among principle- or rule-based theories, the first formulation of Kant's categorical imperative offers a formalizable procedure. I will explore a version of machine ethics along the lines of Kantian formalist ethics, both to suggest what computational structures such a view would require and to see what challenges remain for its successful implementation. In reformulating Kant for the purposes of machine ethics, I will consider three views of how the categorical imperative works: mere consistency, commonsense practical reasoning, and coherency. The first view envisions straightforward deductions of actions from facts. The second view incorporates recent work in nonmonotonic logic and commonsense reasoning. The last view takes ethical deliberation to follow a logic similar to that of belief revision.

Kantian formalist ethics

In *Grounding of the Metaphysics of Morals*,[1] Kant claims that the first formulation of the categorical imperative supplies a procedure for producing ethical rules:

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

Kant tells the moral agent to test each maxim (or plan of action) as though it were a candidate for a universalized rule. Later, he adds that each universalized rule must fit into a system of rules for all persons. In other words, my maxim will be an instance of a rule only if I can will that everyone might act on such a maxim. Further, such a universalized rule must be consistent with other rules generated in a similar manner. Philosophers have interpreted these *universalizability* and *systematicity* conditions as a two-part consistency check on an agent's action plan.

The procedure for deriving duties from maxims—if we are to believe Kant—requires no special moral or intellectual intuition peculiar to humans. For a formalist Kantian, whether a maxim could be a universal rule presents a decision problem that's the same for a human or a machine. Kant himself, 20 years prior to publication of the *Grounding*, sketched an answer to the decision problem that's suggestive of a machine solution:

> 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.[2]

A machine-computable categorical imperative

I don't intend to offer a strict interpretation of Kant's ethics here. Instead, I'll focus on the logic of a machine-computable categorical imperative. Recall that the first formulation is supposed to test maxims. For Kant, maxims are "subjective principles of volition," or plans. In this sense, the categorical imperative 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 that could serve as moral laws—that is, laws holding for any agent. Because we can't stipulate the class of universal moral laws for the machine—this would be human ethics operating through a tool, not machine ethics—the machine might itself construct a theory of ethics by applying the universalization step to individual maxims and then mapping them onto traditional *deontic categories*—namely, forbidden, permissible, obligatory actions—according to the results.

Kant tells the moral agent to test each maxim (or plan of action) as though it were a candidate for a universalized rule. He adds that each universalized rule must fit into a system of rules for all persons.

The first formulation of the categorical imperative demands that the ethical agent act only on maxims that it can universally will. It is somewhat deflating, then, that this formulation gives no more than a necessary condition for ethical action. One simple way to meet this condition would be to universalize each maxim and perform a consistency check. A more efficient method would be to start from scratch and build the theory of forbidden maxims F from the outcomes of consistency checks on possible action plans. The machine would then check whether any prospective maxim m is an element of F. The theory will be finitely axiomatizable if and only if it's identical to the set of consequences of a finite set of axioms. The theory will be complete if and only if, for every maxim m, either it or its negation is in F. If the machine could tell, for any m, whether it's an element of F, then the theory would be decidable. The theory of forbidden maxims (alone) lets the machine refrain from what it ought not do.

This is the optimistic scenario. But how does the machine know what it ought to do? We would need a test that generates the deontic category of obligatory maxims. But a problem arises here if the theory of forbidden maxims is complete. Suppose the categorical imperative assigns the answer "yes" for all forbidden maxims. Two deontic categories still remain for assignment: obligatory and permissible maxims. And, of course, permissible maxims are neither obligatory nor forbidden.

Other problems arise on the formalization level. Consider one that Onora O'Neill discusses.[3] In some cases, we might have a maxim that fails the universalization test because it's overly specific or because of a kind of asymmetry in the predicate. While these are indeed failures, they don't seem to be morally relevant failures. For instance, in the maxim, "I will enslave John," one might not be able to quantify over "John" if he's taken to be a pure existential. In other words, if I want to enslave John because he's a specific person—not just any person—then my maxim won't be applicable to any other object and so won't be universalizable. But the maxim is immoral, of course, not because it's something that I propose to do to John and John only, but because enslaving is wrong. The theory ought to forbid my maxim, but not because of its peculiar specificity.

It's also mistaken to think that slavery is wrong just because of a certain predicative asymmetry in the maxim's universalized form—that I would be willing everyone to be a slave, hence leaving no one to be a slaveholder. Although it's true that one can't be both a slave and a slaveholder, that isn't what makes slaveholding wrong. If the asymmetry were the problem, then maxims such as "I will become a taxi driver" would also fail, on the assumption that we need some people to ride in taxis for others to be employed in driving them.

To address the specificity problem, we must add a condition on a maxim's logical form so that the universalization test will quantify over circumstances, purposes, and agents. If we don't have this restriction, some maxims

might be determinate with respect to either the circumstance or purpose—that is, some might be pure existentials, such as "I will offer *this* prize as a reward." The asymmetry problem is harder to resolve, of course, at least for a machine, because its resolution seems to require some fairly complex semantic ability.

Mere consistency

So now we know that a properly formulated input for testing ethical behavior is a maxim over which circumstances, purposes, and agents are universally quantified. A computer must be able to parse these categories from programmed ontologies, or it must simply accept properly formulated input maxims as having an unambiguous syntax of circumstance, purpose, and agent. 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 categorical imperative's test is an algorithm that alone will determine the classes of obligatory, forbidden, and permissible actions. In other words, the test produces formulas for a deontic logic system.

Now, this deontic logic will include many issues that I must set aside here. Among them are the nature of the logical connectives between circumstances, purposes, and actions; material implication (if-then) is clearly too weak. Another is whether a machine would understand obligation from an agent's perspective—that is, would the machine understand the difference between "I ought to do z" and (merely) "z ought to be the case"? (For more information on this problem, see John Horty's discussion.[4]) So, setting aside these problems, let's suppose that, after the quantification step, the machine can produce universalized maxims that look something like the following (I omit quantifiers here):

1. $(C \text{ and } P) \longrightarrow A$
   A is obligatory for the agent
2. $(C \text{ and } P) \longrightarrow \neg A$
   A is forbidden for the agent
3. $\neg((C \text{ and } P) \longrightarrow A)$ and
   $\neg((C \text{ and } P) \longrightarrow \neg A)$
   A is permissible for the agent

where C represents a circumstance, P represents a purpose, and A represents an action.

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

A major defect in this initial account is apparent if we want the machine to go beyond verifying that a candidate maxim is an instance of one of these three schemata. The categorical imperative doesn't merely perform universal generalization on sentences that are supplied as candidate maxims. Surely, it must test the maxims for contradictions, but the only contradictions that can arise are trivial ones—those inherent in the maxims themselves. This is so even when we take the theory of forbidden maxims to be closed under logical consequence.

A robust version of the test, on the other hand, requires the machine to compare the maxim under consideration with other maxims, principles, and axioms. In other words, the machine must check the maxim's consistency with other facts in the database, some of which will be normative conclusions from previously considered maxims. Obviously, the simple account of mere consistency won't do. It must be buttressed by adding other facts, principles, or maxims, in comparison with which the machine can test the target maxim for contradiction.

> Ethical deliberation conceived as a consistency check on a single universalized maxim is clearly too thin. The main focus for building a Kantian machine should therefore turn to the background theory.

Commonsense practical reasoning

We can buttress Kant's mere consistency test by adding a background theory B, against which the test can have nontrivial results. What would this theory look like? For Kantians, it would depend on the line of interpretation one has for Kant's ethics generally. Many scholars supplement Kant's categorical imperative with normative principles from his other philosophical writings. This way of adding to Kant's pure formulation risks introducing psychological and empirical considerations into practical reasoning. While such considerations seem altogether appropriate to most of us, Kant saw it posing the threat of "heteronomy," thus polluting the categorical imperative's sufficiency to ethical reasoning.

Kant's illustrations of the categorical imperative in the *Grounding* suggest a better alternative. In these illustrations, Kant introduces some *commonsense rules*. For instance, he argues that, because feelings are purposeful and the purpose of the feeling of self-love is self-preservation, it would be wrong to commit suicide out of self-love. He also argues that it's wrong to make false promises because, in general, the practice of giving and accepting promises assumes that promises are kept. Many contemporary Kantians have adopted this suggestion concerning commonsense rules, which they call, variously, postulates of rationality,[5] constraining principles of empirical practical reason,[6] and principles of rational intending.[3] These are presumably nontrivial, nonnormative rules that somehow capture what it is to act with practical reason.

When we build the background theory B with commonsense rules, we get something that is probably closer to ethical deliberation in humans. This move presents difficulties, insofar as we don't have a general formalism for commonsense practical reason (though there are some domain-specific accounts). On the other hand, ethical deliberation conceived as a consistency check on a single universalized maxim is clearly too thin. The main focus for building a Kantian machine should therefore turn to the elements of B; in this way, we might hope to supplement the categorical imperative's test. If this supplementation were successful, we would say that a maxim is unreasonable if it produces a contradiction when we combine it with B. With the proper rules, 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. Logical considerations alone are insufficient to determine whether to include any postulate in B.

Postulates of commonsense practical reason don't share the logic of scientific laws or other universal generalizations. One counterexample is enough to disprove a deductive law, but commonsense postulates must survive the occasional defeat. The postulates of B, then, would require a nonmonotonic theory of practical reasoning.

Nonmonotonic logic attempts to formalize an aspect of intelligence, artificial or human. Nonmonotonic reasoning is quite commonplace. Consider that classical first-order logic is *monotonic*: if you can infer sentence a from a set of premises P, then you can also infer a from any set S that contains P as a subset. Nonmonotonic inference simply denies this condition because the bigger set might contain a formula that "defeats" or disallows the inference to a.

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 that would defeat the inference to "Fritz is affectionate." Let's say we had additional information to the effect that "Fritz is a tiger." At the least, all bets should be off as to whether Fritz is affectionate. An ethics example might be the default rule "Don't kill the innocent." The defeating conditions might be "unless they are attacking under the control of some drug" or "except in a just war," and so on.

While there are different ways to formalize nonmonotonic reasoning, we want to choose a way that will build on the categorical imperative's basic monotonic procedure. We also need a system that extends classical first-order logic and offers the most flexibility, so that we can use the formalism to extend the simple monotonic account of the categorical imperative in the previous section. For these reasons, Reiter's default logic seems to be the best candidate among the approaches developed so far.[7]

In Reiter's default logic, the rule in the example just given becomes

> If Fritz is a cat, *and it is consistent that Fritz is affectionate*, then Fritz is affectionate.

Any number of additional facts can defeat the italic clause, such as "Fritz had a bad day," "Fritz had a bad kittenhood," "Fritz is a person-eater," and so on. Reiter suggests the following symbolization for this default rule:

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

where C is the default's precondition, A is the justification (in this instance), and A is the default conclusion. This is a *normal* default rule because the justification is the same as the conclusion we're allowed to draw. We understand the justification as certifying that no information exists to indicate that the conclusion is false or that Fritz is a special kind of cat—that is, we've learned nothing to convince us that Fritz is not affectionate.

How we use Reiter's default logic depends on the notion of an *extension*, which also appears in other nonmontonic systems. Intuitively, an extension of a theory (T_{ext}) is a set of conclusions of a default theory $T = <W, D>$, where W is a set of facts and D is the set of default rules. We can use a conclusion from the rules in a consistency test, if we can prove the precondition from the set of facts W and if the justifications are consistent with all conclusions of the rules in D. (For further illustrations, see David Poole's work on default logic.[8]) An extension of the theory adds all of those default conclusions consistent with W and its logical consequences, but never adds an untoward fact. Adding the default rules, then, will allow input maxims to contradict the background set of facts and commonsense rules without introducing inconsistency.

While there are different ways to formalize nonmonotonic reasoning, we want to choose a way that will build on the categorical imperative's basic monotonic procedure.

This means the definition of an extension maintains the requirement of nonmonotonicity. Given a set of first-order sentences, we can add the conclusions of default rules without generating conclusions that are inconsistent with the default theory. Default extensions avoid introducing contradictions. Default rules yield to facts; the rules are defeated but not vanquished. In monotonic logic, by contrast, counterexamples vanquish universal laws.

Kant seems to recognize that *defeasible* reasoning—that is, reasoning that displays the property of nonmonotonicity—plays some role in ethical thinking. In this respect, he is far ahead of his time. In the *Grounding*,[1] he refers to a thought process in which 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 the kinds of default rules he might have wanted the background theory to include.

Against suicide

For example, Kant offers the following account of moral deliberation for the person contemplating suicide:

> 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**.

I've added the bold font to what I take to be nonmonotonic reasoning. The default rule concerns the function or purpose of self-love, premise 3 in the reconstructed argument that 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.

The normal default rule allows self-preservation from the precondition of self-love, provided that self-preservation is consistent with other facts and default-rule conclusions. But self-preservation is no universal duty 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. Lacking those defeating conditions, and provided that the agent satisfies the antecedent conditions, the universalized maxim plus the default rule seems to yield the contradiction that the categorical imperative needs.

What happens when two default rules yield incompatible conclusions? Suppose we have two default rules in the theory:

- Suicide is self-destruction.
- Martyrdom is honorable.

Here, we could face the problem of multiple extensions: one rule tells us one thing, and the other allows us to infer the opposite. (A standard example of a harder case is "Republicans are hawks," "Quakers are pacifists," and the additional fact that "Nixon is a Republican Quaker.") This problem could arise in machine ethics, in which case we would need some procedure for specifying rule priorities.

Against false-promising

A second example of nonmonotonic reasoning appears in Kant's account of 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**.[1]

Again, I've added the bold font to highlight nonmonotonic reasoning. The traditional criticism of this illustration is that promising and borrowing would *not* in fact be impossible if false promising became a universal rule in the closely defined circumstance of need. Such a condition would only engender extreme caution in lending and an insistence on collateral.

I don't believe this objection holds, however, because it misses the defeasible nature of both promising and lending. The institution of promising depends on two default rules—one for the debtor and one for the creditor—that promises are believed and 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 isn't 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—a universal generalization, "All promises are kept or promising is destroyed"—doesn't properly interpret the institution. The actual institution of promising depends as much on *surviving* a defeating instance as it does on the prevalence of nondefeat. So a nonmonotonic interpretation of the illustration makes sense of the practice, while the monotonic interpretation does not.

Difficulties for the nonmonotonic approach

The nonmonotonic approach to deontological machine ethics involves one serious problem. Nonmonotonic inference fails a requirement met by classical first-order logic:

> We need a background theory of commonsense reasoning for the categorical imperative test to give nontrivial results. Monotonic logic doesn't entirely capture commonsense reasoning.

semidecidability of set membership. Recall the earlier characterization of the categorical imperative as asking whether a candidate maxim is forbidden. Because questions in nonmonotonic logic aren't semidecidable, it's not even the case that the nonmonotonically enhanced categorical imperative is guaranteed to answer "yes" to the question, even when the maxim is in fact forbidden. Of course, by the definition of semidecidability, it's also not guaranteed to answer "no."

The obvious question here is: What good is the nonmonotonic categorical imperative? Let me summarize the general predicament. The nonmonotonic account of Kant's illustrations interprets the ethical deliberation procedure better than anything offered by monotonic logic. We need a background theory of commonsense reasoning for the categorical imperative test to give nontrivial results. Monotonic logic doesn't entirely capture commonsense reasoning. Kant himself, when he does provide clues as to the "buttressing" principles he assumes, gives us rules that can only make sense if they're default rules. But this revised interpretation still fails an important *formal* requirement for machine ethics: semidecidability.

Coherency

In the third candidate for the logic of machine ethics, ethical deliberation involves the construction of a *coherent system* of maxims—a system that accepts any minimal set of consistent maxims as the background for comparing any current maxim for consistency. Kant also suggests this view, so it will help if we return to his illustrations of the categorical imperative in the *Grounding*.[1]

These illustrations concern the duties to develop your own talents and to give to others in need. One reading of these illustrations might go as follows: a maxim allowing your talents to rust conflicts with what every rational being wills, according to Kant—namely, the development of your talents. And if you want help from others when you're in need, you must agree to help others when they're in need. What these cases share is the prohibition against acting on a maxim that is incoherent, given a minimal set (perhaps singleton set) of other maxims. The other maxims provide the coherency constraint but aren't privileged by stipulation; nor are they conclusions from nonmonotonic reasoning. They are your own maxims. Presumably, a machine could build such a database of its own maxims.

Let's consider the categorical imperative's procedure as a kind of bottom-up construction. Ethical deliberation, in this view, should be like building a theory, where the theory's sentences are your own maxims plus any of their consequences. Call this theory G. The theory also has two rules: R-in and R-out. For any maxim m_i, in the set of maxims M on which the machine is now prepared to act, R-in says that m_i is allowed in M if and only if m_i and G are consistent.

What about maxims the machine has acted on in the past that subsequently turned out to be impermissible? Handling such incoherencies is analogous to the belief-revision problems that Peter Gärdenfors explored.[9] If we allow the "impermissible" maxims to remain in G, the set of sentences will automatically be inconsistent; hence, the coherency constraint breaks down. Surely Kant doesn't insist on past moral perfection as a condition for reasoning about right action in the present.

We can now describe a rule (R-out) for excluding maxims that would maintain the

set's inconsistency. There's 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 machine, but how the machine learns that some maxim was wrong remains a mystery. We can call the procedure an update, but that doesn't indicate how the machine would update itself. Because this seems to be crucial to ethical deliberation, this model still doesn't yield an ethical machine.

Another interesting aspect of G that poses a difficulty for a Kantian machine is the limiting case where m_1 is the only member of M. We might call this the case of the moral infant. G must allow a first maxim to enter by *R*-in because, by hypothesis, G is empty and so it's 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 can't 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 *n*th 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? Of course, good practical grounds exist for a moral agent to hold to the set of maxims already accumulated. Moreover, we might think that no typical moral agents are moral infants because everyone has, at any given time, an established set of maxims. But is it not true that all potentially ethical machines will be moral infants, at some point in time? To construe Kant's test as a way to "build" a set of maxims, we must establish priority rules for accepting each additional maxim. We must have what Gärdenfors calls an *epistemic commitment function*,[9] though ours will be specific to moral epistemology. This is a species of the more general problem with antifoundationalist epistemology; not all knowledge can depend on other knowledge.

The problem of the moral infant shows that a Kantian formalism in the constructivist or "bottom-up" tradition can't build a coherent moral theory from nothing. A deontological theory must give reasons why the machine shouldn't throw out an entire collection of maxims to allow entry of one otherwise incoherent maxim, m_n. In terms of human ethics, a Kantian theory must tell agents who've compiled good moral character why they can't now defeat all of those prior maxims and turn to a life of vice. I think a Kantian could give many good reasons, but not the ones that a bottom-up constructivist theory offers.

I've suggested three accounts, according to which we might conceive of a deontological ethical machine. Each account has its challenges—triviality, asymmetry, excessive specificity, lack of semidecidability, and lack of priority for maxims, to repeat those I've described here. Although these problems seem difficult to surmount, they are similar to problems in human attempts to engage in practical reasoning. Ethicists have explicated these problems for centuries, yet few of us have given up on the general view that our action plans include formal properties that mark them as right or wrong. Perhaps work on the logic of machine ethics will clarify the human challenge as well. ■

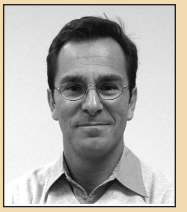
Acknowledgments

I would like to thank Colin Allen, Fred Adams, Nicholas Asher, Amit Hagar, and several anonymous reviewers for critical comments on this article.

References

1. I. Kant, *Grounding for the Metaphysics of Morals*, translated by J. Ellington, Hackett, 1981.
2. I. Kant, *Bemerkungen in den "Beobachtungen über das Gefühl des Schönen und Erhabenen"* [Unpublished Notes on "Observations on the Feeling of the Beautiful and the Sublime"], Felix-Meiner Verlag, 1991 (in German, translated by the author).
3. O. O'Neill, *Constructions of Reason*, Cambridge University Press, 1989.
4. J. Horty, *Agency and Deontic Logic*, Oxford Univ. Press, 2001.
5. J. Silber, "Procedural Formalism in Kant's Ethics," *Review of Metaphysics*, vol. 28, 1974, pp. 197–236.
6. J. Rawls, "Kantian Constructivism in Moral Theory," *J. Philosophy*, vol. 77, no. 9, 1980, pp. 515–572.
7. R. Reiter, "A Logic for Default Reasoning," *Artificial Intelligence*, vol. 13, 1980, pp. 81–132.
8. D. Poole, "Default Logic," *Handbook of Logic in Artificial Intelligence and Logic Programming*, D. Gabbay, C. Hogger, and J. Robinson, eds., Oxford, Univ. Press, 1994.
9. P. Gärdenfors, *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, MIT Press, 1988.

The Author



Thomas M. Powers is an assistant professor of philosophy at the University of Delaware and a faculty research fellow at the Delaware Biotechnology Institute. He received his PhD in philosophy from the University of Texas at Austin. His research focuses on the intersection of ethics, science, and technology. He's a member of the AAAI, the American Philosophical Association, the International Society for Ethics and Information Technology, and the Association for Practical and Professional Ethics. Contact him at the Dept. of Philosophy, Univ. of Delaware, Newark, DE 19716; tpowers@udel.edu.