Toward Non-Intuition-Based Machine Ethics

J. N. Hooker and Tae Wan Kim
Carnegie Mellon University, Pittsburgh, USA
twkim,jh38@andrew.cmu.edu

Abstract

We propose a deontological approach to machine ethics that avoids some weaknesses of an intuition-based system, such as that of Anderson and Anderson. In particular, it has no need to deal with conflicting intuitions, and it yields a more satisfactory account of when autonomy should be respected. We begin with a "dual standpoint" theory of action that regards actions as grounded in reasons and therefore as having a conditional form that is suited to machine instructions. We then derive ethical principles based on formal properties that the reasons must exhibit to be coherent, and formulate the principles using quantified modal logic. We conclude that deontology not only provides a more satisfactory basis for machine ethics but endows the machine with an ability to explain its actions, thus contributing to transparency in AI.

Introduction

No one has done more to develop machine ethics than Anderson and Anderson (2007, 2010, 2011, 2014, 2015a, 2015b; with Armen 2006; with Berenz 2017). Throughout the debate between case- and principle-based approaches (Wallach, Allen, and Smit 2008), Anderson and Anderson (hereafter, A&A) show that it is better for machine ethics to be principle-based (cf. Guarini 2011). In particular, they show that any adequate machine ethics should have non-consequentialist elements to respect the dignity of persons.

We build upon A&A's work, but critically. They computerize W. D. Ross's (1930) *prima facie* duty approach for the context of health-care scenarios (e.g., GENETH, EthEl). We argue that their *prima facie* duty approach is inadequate for machine ethics for two reasons: (i) it relies on human moral intuition; (ii) its treatment of autonomy is inadequate.

We propose a non-intuition-based machine ethics that rests, instead, on deontology. We show that ethical rules for machines can be derived from first principles and stated with a reasonable degree of rigor. We formulate these rules in the idiom of quantified modal logic, which is well suited to formalize the role of rational belief in deontological ethics. We show that this approach can avoid some of the pitfalls of relying on moral intuition and yield a more adequate analysis of autonomy.

We begin with a critique of A&A's approach. We then sketch a "dual standpoint" theory of action that regards actions as necessarily grounded in reasons and therefore having a conditional form, which is well suited to formulating rules for machines. We then derive ethical principles based on formal properties that the reasons for an action must have in order to be coherent, and formulate the principles using quantified modal logic. At this point we can show how a healthcare example central to A&A's discussion is more adequately treated with deontological ethics. The paper concludes by pointing out that the same approach installs an ability for the machine to explain its actions, thus contributing to transparency in AI.

A&A's Prima Facie Duties Approach

A&A use inductive logic programming to discover a decision principle for a specific domain. To make this concrete, consider one of A&A's example scenarios:

> A doctor has prescribed a medication that should be taken at a particular time in order for the patient to receive a small benefit (i.e., the patient will be more comfortable); but, when reminded, the patient doesn't want to take it at that time.

The normative question is:

> Should the system notify the overseer that the patient won't take the medication at the prescribed time or not?

The system has two options: *Notify* and *Don't notify*. Relying upon the work of biomedical ethicists Buchanan and Brock (1990), A&A assume that the correct answer is *Don't notify*. To analyze the rationale, drawing upon the work of biomedical ethicists Beauchamp and Childress (1979), A&A assume that three *prima facie* duties—nonmaleficence, beneficence, and autonomy—are relevant to the scenario. A&A associate either option (*Notify* or *Don't notify*) with an ordered triple (v_1, v_2, v_3) that indicates the degree to which each of the three duties is satisfied.

Harm does not result from either option. *Notify* achieves beneficence but violates autonomy, while *Don't notify* sacrifices beneficence while respecting autonomy. Suppose we set the value of beneficence at $+1$ and its absence at -1, and we set the value of respecting autonomy at $+2$ and violating it at -2. Then the two options are associated with ordered triples as follows:

Notify:	$(0, 1, -2)$
Don't notify:	$(0, -1, 2)$

Repeating the process for other variations that cover all the possible cases, the system inductively seeks an equilibrium that coherently covers all the choices that A&A assume to be as correct, based on Buchanan and Brock (1990). As a result, the decision principle that A&A's system discovers in the above healthcare context is, "A healthcare worker should challenge a patient's decision if it isn't fully autonomous and there's either any violation of non-maleficence or a severe violation of beneficence."

When a duty of beneficence and a duty of autonomy conflict as in the case above, A&A's system relies upon the moral intuition of ethicists, as advocated by the intuitionist W. D. Ross (1930). Ross argued that ethics consists of various duties that sometimes conflict with each other, and that when there is conflict, we should rely on the moral intuitions of "well-educated people." He writes,

> [M]oral convictions of thoughtful and well-educated people are the data of ethics just as sense-perceptions are the data of a natural science (1930: 41)

In the same vein, A&A (2011) write,

> We used ethicists' intuitions to tell us the degree of satisfaction/violation of the assumed duties within the range stipulated, and which actions would be preferable, in enough specific cases from which a machine-learning procedure arrived at a general principle (479)... We believe that there is an expertise that comes from thinking long and deeply about ethical matters. Ordinary human beings are not likely to be the best judges of how one should behave in ethical dilemmas (482).

Assuming the correctness of the ethicist's intuition—that *Don't notify* is the right choice—A&A's system analyzes the rationale. That is, the system is trained to discover a coherent set of principles using the ethicists' moral intuition as training data.

Problems of Moral Intuition in Machine Ethics

Relying on moral intuition is often problematic. Intuition may be an important part of ethical reasoning (see, e.g., reflective equilibrium in Rawls 1971) but is not itself an argument (Dennett 2013). Furthermore, experimental philosophers show that moral intuitions are not as consistent as we think. For example, they are susceptible to morally irrelevant situational cues (e.g., Alexander 2012, Appiah 2008, Sinnott-Armstrong 2006). A&A might respond that their system relies on the intuition of a professional ethics researcher. But evidence shows that professional intuitions are not significantly different from those of ordinary people (for reviews, see Schwitzgebel and Rust 2016).[1]

Additionally, a major rationale behind developing machine ethics is inconsistent with reliance on human intuition. For example, a prime motivation for developing autonomous vehicles is that human error is a leading cause of accidents, and autonomous vehicles minimize human involvement. Thus if human moral intuitions are not reliable, machine ethics should be developed so as to avoid human moral intuition.

Finally, A&A's system is helpless in situations about which professional ethicists' intuitions do not have consensus. In response, A&A argue that machines should not be allowed to make a choice for such cases. But machines may face scenarios in which they must make choices (not making a choice is itself a choice). And, as A&A often emphasize, one contribution of machine ethics should be to make breakthroughs in dilemmas that human ethicists cannot resolve. We therefore develop a non-intuition-based approach.

A&A's Treatment of Autonomy

In their earlier works (2007, 2006 with Armen), A&A considered the adequacy of hedonic utilitarianism for machine ethics because the theory is straightforward to codify. The machine need only maximize expected net pleasure. But soon they turned to the *prima facie* duty approach because of standard problems in utilitarianism: it demands sacrificing the good of one for the good of many, which means that respect for an individual's autonomy is not guaranteed.

We agree with A&A that machine ethics must possess deontological elements that protect a person's dignity. But we believe that the way A&A computerize autonomy fails to fulfill its expected role. It works in the above scenario because there is only one patient, and the patient is the only beneficiary. But imagine a case in which violating one person's autonomy maximizes benefit for ten thousand others. For example, a company might send a manager unjustly accused of fraud to prison (by falsely testifying against her) in order to satisfy the media and boost share prices. To avoid endorsing this outcome, A&A's system must say that the value of autonomy violation in this case (and relevantly similar cases) is $-10{,}001$ to counterbalance a total beneficence score of 10,000. Yet if there is only one stockholder, the system would presumably set the value of autonomy violation at -2 because the value of beneficence is only 1. We do not see why the value of the manager's autonomy should differ dramatically, depending upon the number of parties who stand to benefit. This problem occurs because there is no adequate notion of autonomy in A&A's system.

Deontological Ethics for Machines

The vagaries of intuition-based ethics can largely be avoided by a rigorous development of deontological ethics for machines. "Deontological" literally means duty-oriented, but it is normally interpreted as referring to rule-oriented ethics, which encodes obligation as rules of conduct. While deontology is inseparable from the name of Immanuel Kant, it need not be bound to Kant's historical theory. The basic reasoning process one finds in Kantian ethics can give rise to precise rules that are grounded in first principles.

[1] The cited paper does not survey works that directly study the moral intuition of ethicists, but the consistency between their ethical beliefs and behaviors. If it is plausible to believe that humans often use intuitions to guide behavior, the research implies that ethicists' moral intuitions are not more reliable than those of ordinary people.

The key is to recognize that actions are necessarily *based on reasons* (Anscombe 1957; Davidson 1963). While all behavior is determined by physical and biological causes, a "dual standpoint" theory allows one to distinguish action from mere behavior by virtue of the fact that it has a *second kind of explanation*, namely the agent's reasons for undertaking the action. Thus an action can be viewed from two standpoints, namely as a result of physical causes, and as the conclusion of a reasoning process.[2]

Because actions are always based on reasons, they have a conditional form: "If such-and-such reasons apply, then perform such-and-such an action." Conditional rules of this kind are naturally suited to be programmed into a machine. We can also derive necessary conditions for an *ethical* rule by requiring that the reasons satisfy certain formal consistency properties, as is the tradition in deontological ethics.

For present purposes we do not regard machines as agents that exercise autonomy in the full sense necessary to derive ethical obligations for the machines themselves. They are seen as autonomous only in the limited sense that they operate according to internal rules and are not under constant human control. Rather, the human programmer is the agent, and rules encoded into the machine must be rules that the human can ethically inject into a machine.

Action Plans

Because actions necessarily have conditional form, we will refer to them as *action plans*. As an example, suppose I walk into a department store and see a display of watches. The watches are in an open case, and as I look around, I see that there is nothing to prevent me from stealing one (no security guards, no surveillance cameras, etc.). So I steal a watch. Let's suppose my action plan is a conditional statement:[3]

$$C_1 \wedge C_2 \Rightarrow A_1$$

where the antecedents C_1, C_2 and consequent A_1 are interpreted

$C_1 =$ "I would like to have a new watch."
$C_2 =$ "I can get away with stealing one."
$A_1 =$ "I will now steal the watch."

[2]The phrase "dual standpoint" derives from Kant's statement that "the concept of a world of understanding is therefore only a *standpoint* that reason sees itself constrained to take outside of appearances *in order to think of itself as practical*" ("Der Begriff einer Verstandeswelt is also nur ein *Standpunkt*, den die Vernunft sich genöthigt sieht, außer den Erscheinungen zu nehmen, *um sich selbst als praktisch zu denken*") (Kant 1785, page 458). In other words, to see oneself as taking action (in Kantian language, to think of oneself as "practical"), one must interpret oneself as existing outside the natural realm of cause and effect. Or to use more modern language, one must be able to give one's behavior a second kind of explanation that is based on reasons one adduces for it, rather than on cause and effect. This idea eventually evolved into the dual standpoint theories of recent decades (Nagel 1986; Korsgaard 1996; Bilgrami 1996), which have some parallels to the theory adopted here.

[3]Properly speaking, an action plan is *expressed* or *denoted* by a conditional statement, but we will blur the use/mention distinction to simplify exposition. To avoid excessive parentheses, we will write $(C_1 \wedge C_2) \Rightarrow A_1$ as $C_1 \wedge C_2 \Rightarrow A_1$.

The symbol $\Rightarrow$ means that I take the conditions C_1, C_2 to be sufficient reason to carry out action A_1 when they are satisfied. Thus C_1, C_2 are not psychological causes or motivations for taking action A_1, but conditions that I regard as sufficient reason for taking action A_1 as part of my reasoning process.

I must regard the antecedents of my action plan as jointly *sufficient* and individually *necessary* for the action. Sufficiency means that I have decided to take the action whenever the reasons apply, since otherwise they are not a complete list of reasons. Necessity means that I will not necessarily take the action if any of the conditions are missing. I may, of course, view additional conditions as necessary to justify my theft: I believe no one at the shop is likely to be fired as a result of my theft, I do not expect to feel remorse, etc.

We also require that an action plan be *maximal*; that is, as general as possible while remaining a valid description of the agent's reasoning. Suppose I am an ambulance driver and find that, due to heavy traffic, I will be late for an appointment with my boss. I therefore drive my ambulance to the appointment while using siren and lights, even though there is no medical emergency. I have what appears to be an action plan $C_4 \wedge C_5 \wedge C_6 \Rightarrow A_2$, where

$C_4 =$ "I am late for an appointment with my boss."
$C_5 =$ "The traffic is heavy enough to make me late unless I use the siren and lights."
$C_6 =$ "I can get away with using siren and lights when there is no medical emergency."
$A_2 =$ "I will use siren and lights."

However, this is probably not a true account of my reasons. Suppose that in another scenario, I must pick up my kids at day care, and I am running late because I had problems starting the engine. Why wouldn't I use the siren and lights in this case? If I have no particular reason for doing so in one case and not the other, it is evident that my true reasons are more general:

$C_4' =$ "It is really important to be on time."
$C_5' =$ "I will be late unless I use the sirens and light."

as well as C_6. So my action plan is actually $C_4' \wedge C_5' \wedge C_6 \Rightarrow A_2$. We can express this formally as follows:

> If $\mathcal{C} \Rightarrow A$ is an action plan, and $\mathcal{C}' \rightarrow \mathcal{C}$ but $\mathcal{C} \not\rightarrow \mathcal{C}'$, then $\mathcal{C}' \Rightarrow A$ is not an action plan.

where $\mathcal{C}$ and $\mathcal{C}'$ are conjunctions of conditions, and $\mathcal{C}' \rightarrow \mathcal{C}$ means that $\mathcal{C}'$ implies $\mathcal{C}$.

From Action Theory to Ethics

The next step is to derive necessary conditions for ethical action plans that are based on the consistency of the rationales in the plans. This can be accomplished by appealing to the universality of reason: the validity of one's reasoning process should not depend on who one is. If I take certain reasons to justify my action, rationality requires me to take them as justifying this action for anyone to whom the reasons apply.

For example, suppose that I lie simply because it is convenient to deceive someone. Then when I decide to lie for

this reason, I decide that everyone should lie whenever deception is convenient. Every choice of action for myself is a choice for all agent. This leads to the famous *generalization principle*, which is perhaps best stated as follows: I must be rational in believing that the reasons for my action are consistent with the assumption that everyone with the same reasons takes the same action. An action plan that satisfies this principle is *generalizable*. Onora O'Neill (2014) provides an excellent reconstruction of thought along this line.

Thus if I lie because it is convenient to deceive someone, I am adopting this as a policy for everyone. Yet I am rationally constrained to believe that if everyone in fact lied when deception is convenient, no one would believe the lies, and no one would be deceived. My reasons for lying would no longer justify lying. So the reasons behind my decision to lie are ungeneralizable and therefore self-contradictory.

Formalizing the Generalization Principle

We can formalize the generalization principle to a certain degree by introducing elements of quantified modal logic. First, we regard a condition C_i as a predicate that applies to an agent. Thus $C_i(a)$ states that agent a satisfies condition C_i. It is also convenient to let $\mathcal{C}(a)$ serve as shorthand for the conjunction $\bigwedge_{C\in\mathcal{C}} C(a)$. If agent a adopts the action plan $\mathcal{C} \Rightarrow A$, we write $\mathcal{C}(a) \Rightarrow A(a)$.[4]

Consider again the theft of the watch, which we suppose is carried out by agent a and represented by the action plan $C_1(a) \wedge C_2(a) \Rightarrow A(a)$. One of the reasons for the action is $C_2(a)$: agent a can get away with the theft. However, there are many customers entering the shop who would like a new watch and could steal one with impunity. Rationality constrains agent a to believe that if everyone were to adopt this action plan, reason $C_2(a)$ would no longer apply. The shop would crack down by displaying the watches under glass, installing security systems, and so forth.

We can write this more formally by borrowing the operators $\Box$ and $\Diamond$ from modal logic. We define $\Box S$ to mean that the agent is rationally constrained to believe proposition S; that is, it is irrational for the agent to deny S. We define $\Diamond S$ to mean $\neg\Box\neg S$, or it is not irrational for the agent to believe S. We will frequently express $\Diamond S$ by saying simply that the agent can rationally believe S. Note that these definitions differ somewhat from those normally used in epistemic and doxastic logics.

To formulate generalizability, we adopt the notation $P(S)$ to mean that it is physically possible for proposition S to be true. The action plan $C_1(a) \wedge C_2(a) \Rightarrow A_1(a)$ is generalizable only if a can rationally believe that it is possible for a to carry out the theft when everyone adopts this plan:

$$\Diamond P\Big(C_1(a) \wedge C_2(a) \wedge A_1(a) \wedge \forall x\big(C_1(x) \wedge C_2(x) \Rightarrow A_1(x)\big)\Big)$$

Since a cannot rationally believe this, the action plan is unethical. This leads to the rule

Action plan $\mathcal{C}(a) \Rightarrow A(a)$ is generalizable only if

$$\Diamond P\Big(\mathcal{C}(a) \wedge A(a) \wedge \forall x\big(\mathcal{C}(x) \Rightarrow A(x)\big)\Big)$$

We can now see the importance of requiring action plans to be maximal. In the case of an ambulance driver a, the (maximal) action plan $\{C_4'(a), C_5'(a), C_6(a)\} \Rightarrow A_2(a)$ violates the generalization principle because

$$\neg\Diamond P\Big(C_4'(a) \wedge C_5'(a) \wedge C_6(a) \wedge A_2(a) \wedge \forall x\big(C_4'(x) \wedge C_5'(x) \wedge C_6(x) \Rightarrow A_2(x)\big)\Big)$$

It is not reasonable for a to believe that a can get away with his mischief if all ambulance drivers used the siren and lights whenever they are in a hurry to meet an important engagement. The city would crack down on the practice and carefully monitor drivers. However, the narrower rule $C_3(a) \wedge C_4(a) \wedge C_5(a) \Rightarrow A_2$ is generalizable because circumstances C_3 and C_4 rarely occur, and a can rationally believe that a could get away with it if ambulance drivers always misbehaved in these circumstances. So if the narrower rule were treated as an action plan, we would have to conclude that misusing the ambulance is generalizable.

Respecting Autonomy

An agent violates autonomy when one of its action plans interferes with action plans of one or more other agents. Generally, this occurs when there is coercion, bodily injury that impairs another agent, suppression of another's rational faculties, or death.

It is no violation of autonomy, however, to interfere with behavior that has no coherent rationale, because in this case the behavior is not action. This could occur if the agent simply has not formulated a clear rationale, or the agent's rationale violates one of the conditions for ethical choice. For example, it is no violation of autonomy for me to prevent you from stealing a bicycle from a bicycle rack. This type of coercion may be unethical for other reasons, but it is not a violation of autonomy.

The conditional form of action plans also helps to distinguish ethical from unethical interference. Suppose, for example, that you decide to cross the street to catch a bus as soon as no cars are coming. You begin to cross, but I grab you by the arm and pull you off the street. This is an obvious violation of autonomy, because my action plan interferes with yours. On the other hand, suppose there is a car coming, I shout a warning that you cannot hear, and I then pull you out of the path of the car. I prevent you from crossing the street, but there is no violation of autonomy, because my action does not interfere with your action plan.

Finally, the conditional form accommodates the principle that no violation of autonomy occurs if there is informed consent to interfere. If b has given informed consent to be blocked from action $A(b)$ under condition $C(b)$, then any of b's rational action plans that result in action $A(b)$ will have the form $\neg C(b) \wedge \mathcal{C}(b) \Rightarrow A(b)$. Thus another agent b who has obtained informed consent will not violate b's autonomy by interfering with $A(b)$ under condition $C(b)$. This will be illustrated in our discussion of the medical example.

[4]The constant a may be absent from some or all of the conditions, and other constants may appear to represent agents affected by the action.

Formally, an action plan $\mathcal{C}_1(a) \Rightarrow A_1(a)$ interferes with an action plan $\mathcal{C}_2(b) \Rightarrow A_2(b)$ when $\mathcal{C}_1(a) \wedge \mathcal{C}_2(b)$ is physically possible and $A_1(a)$ interferes with $A_2(b)$. We denote interference with the special notation $A_1(a) \hookrightarrow \neg A_2(b)$ to reflect the fact that the contrapositive does not hold. Thus we do not have $A_2(b) \hookrightarrow A_1(a)$, because we do not wish to say that my crossing the street interferes with your pulling me off the street. We do not further analyze the concept of interference here but assume that it is clear enough for practical purposes.

In general, we say that $C(a) \Rightarrow A(a)$ interferes with a set

$$\{\mathcal{C}_i(a_i) \Rightarrow A_i(a_i) \mid i \in I\} \tag{1}$$

of action plans when $a \neq a_i$ for all $i \in I$ and

$$P\Big(C(a) \wedge \bigwedge_{i\in I} C_i(a_i)\Big) \wedge \Big(A(a) \hookrightarrow \neg \bigwedge_{i\in I} A_i(a_i)\Big)$$

A *joint autonomy principle* can now be formulated. An action plan $C(a) \Rightarrow A(a)$ violates the joint autonomy of a set $\{a_i \mid i \in I\}$ of agents when the agents have a set (1) of action plans such that a is rationally constrained to believe $C(a) \Rightarrow A(a)$ interferes with these action plans. More precisely,

$$\begin{aligned}\Box\Big(P\big(C(a) \wedge \bigwedge_{i\in I} C_i(a_i)\big)\Big) \\ \wedge \Box\Big(A(a) \hookrightarrow \neg \bigwedge_{i\in I} A_i(a_i)\Big)\end{aligned} \tag{2}$$

Your action plan in the example is something like $C_7(b) \wedge \neg C_8(b) \Rightarrow A_3(b)$, where you are agent b and

$C_7(b)$ = "Agent b wishes to go to the bus stop across the street,"

$C_8(b)$ = "There are cars approaching that would endanger agent b," and

$A_3(b)$ = "Agent b will cross the street now."

In the scenario in which I violate your autonomy, I have the action plan $C_7(b) \wedge \neg C_8(b) \wedge A_3(b) \Rightarrow A_4(a, b)$, where I am agent a and

$A_4(a, b)$ = "Agent a will pull agent b off the street."

I violate your autonomy because

$$\Box P\big(C_7(b) \wedge \neg C_8(b) \wedge A_3(b)\big) \wedge \Box\big(A_4(a, b) \hookrightarrow \neg A_3(b)\big)$$

In the scenario in which I pull you out of the path of a car, my action plan is $C_8(b) \wedge A_3(b) \Rightarrow A_4(a, b)$. Even though we still have the interference $A_4(a, b) \hookrightarrow \neg A_3(b)$, I do not violate your autonomy because

$$\neg\Box P\big(C_7(b) \wedge \neg C_8(b) \wedge C_8(b) \wedge A_3(b)\big)$$

In fact, the conjunction is logically as well as physically impossible.

The deontological argument for the joint autonomy principle stems again from the fact that an agent legislates for all agents, due to the universality of reason. A set of action rules for agents in general can be rational only if they are compatible with each other, meaning that none can interfere with another. Thus if an agent adopts an action plan that interferes with the action plans of others, the agent introduces incompatability into the set of action rules for agents in general. This is irrational and therefore unethical.

Utilitarian Principle

Utilitarianism is normally conceived as a consequentialist theory but can be formulated deontologically as well. That is, rather than judging an act by the net expected utility it actually creates, the principle can require that an agent take actions that it can rationally believe maximize what it regards as utility. Utility is defined as a state of affairs that the agent regards as an end in itself rather than a means to some other end, such as happiness.

We might formalize the utilitarian principle as follows. We suppose there is a utility function $u(A(a), \mathcal{C}(a))$ that measures the total net expected utility of action $A(a)$ under conditions $\mathcal{C}(a)$. Then an action plan $\mathcal{C}(a) \Rightarrow A(a)$ satisfies the *utilitarian principle* only if agent a can rationally believe that it creates at least as much utility as any ethical action plan that is available under the same circumstances. That is, $\mathcal{C}(a) \Rightarrow A(a)$ satisfies the utilitarian principle only if

$$\Diamond\forall A'\Big(u(A(a), \mathcal{C}(a)) \geq u(A'(a), \mathcal{C}(a))\Big)$$

where A' ranges over all actions that are available under conditions $\mathcal{C}(a)$ and satisfy the generalization and joint autonomy principles. Note that we are now quantifying over predicates and have therefore moved into second-order logic.

The Medical Example

We can now return to the medical example posed by A&A and ask how a machine should be programmed to respond when a patient refuses to take medication at the prescribed time, based on the deontological criteria developed above. We will again suppose that there are two options, *Notify* and *Don't notify*, meaning that the machine will or will not notify the supervisor of the patient's refusal.

The scenario is set up in such a way neither option causes harm to the patient. Since taking the medication on time creates some benefit (it makes the patient "more comfortable"), we will assume that *Notify* results in greater expected utility that *Don't notify*.

There is no clear reason to believe that either option violates the generalization principle. In any case, we will assume there is no violation, since the main focus of our analysis in this scenario is on autonomy.

We cannot accept A&A's assumption that notifying the supervisor violates patient autonomy. They base this judgment on Buchanan and Brock's book (1990), but on examination of the book, we can find no justification for it other than the fact that it reflects the expert opinion of these authors.

Indeed, it is clear that notifying the supervisor does not, in and of itself, interfere with any of the patient's action plans. The patient may *prefer* that the supervisor not be notified, but a preference is not an action plan. The patient may have an action plan of not personally notifying the supervisor, but there is no interference with this plan. There could be a violation of autonomy if notification triggers some restriction of privileges the patient plans to exercise, but this is not part of the scenario.

This outcome illustrates the risk of relying on moral intuition. Notification may "go against patient wishes" in some colloquial sense, perhaps in the sense that the patient doesn't *want* the supervisor to know. Yet notification is not a violation of autonomy, because autonomy is assessed only with respect to the patient's *actions*.

For illustrative purposes, however, we will modify the scenario so that notification does in fact result in a restriction of privileges the patient wishes to exercise. Even so, there is no violation of autonomy if the patient has given informed consent. That is, if the patient entered the institution with the understanding that refusal to follow medical advice could result in restriction of certain privileges.

A rule that respects autonomy is therefore

$$C_9(b) \wedge \neg C_{10}(b) \Rightarrow A_5(a,b) \tag{3}$$

where

$C_9(b)$ = "Patient b has given informed consent to notification,"
$C_{10}(b)$ = "Patient b takes medication at the prescribed time," and
$A_5(a,b)$ = "The system a will inform a supervisor that patient b failed to take medication at the prescribed time."

We can verify as follows that policy (3) respects autonomy. Two relevant and coherent action plans are available to patients in an institution that adopts policy (3), assuming they are aware of the policy:

$$C_{10}(b) \wedge C_{11}(b) \Rightarrow A_6(b) \tag{4}$$
$$\neg C_9(b) \wedge \neg C_{10}(b) \wedge C_{11}(b) \Rightarrow A_6(b) \tag{5}$$

where

$C_{11}(b)$ = "Patient b wants to exercise privileges that are revoked after failure to take medication at the prescribed time," and
$A_6(b)$ = "Patient b will exercise privileges that are revoked after failure to take medication at the prescribed time."

Action plan (4) is relevant when the patient has taken medication as prescribed. It is coherent because the patient knows that, in this case, he/she can exercise privileges regardless of prior consent. Action plan (5) is relevant if the patient has refused to take medication as prescribed. As noted earlier, a rational action plan must include the condition $\neg C_9(b)$.

Although we are rationally constrained to believe that $A_5(a,b) \hookrightarrow \neg A_6(b)$, policy (3) interferes with action plan (4) if and only if (2) holds. That is, both of the following must hold:

$$\Box P\big(C_9(b) \wedge \neg C_{10}(b) \wedge C_{10}(b) \wedge C_{11}(b)\big) \tag{6}$$
$$\Box\big(A_5(a,b) \hookrightarrow \neg A_6(b)\big) \tag{7}$$

While (7) is satisfied, (6) is clearly violated due to the logical contradiction between $\neg C_{10}(b)$ and $C_{10}(b)$, and so there is no violation of autonomy. Similarly, (3) interferes with action plan (5) if and only if (7) and

$$\Box P\big(C_9(b) \wedge \neg C_9(b) \wedge \neg C_{10}(b) \wedge C_{11}(b)\big)$$

The latter is violated, and policy (3) therefore respects autonomy.

On the other hand, a policy of notifying without informed consent is

$$\neg C_{10}(b) \Rightarrow A_5(a,b)$$

which interferes with action plan (5). The policy is therefore unethical if any patients have this action plan.

Naturally, informed consent must be both informed and consensual. The institution staff must satisfy themselves that the patient understands the consequences of notification if the patient fails to take medication as prescribed. This might be accomplished by discussing the issue with the patient rather than simply giving a patient a sheaf of papers to sign. Also the patient must have realistic options other than giving consent, so that there is no coercion (and violation of autonomy) in process of obtaining consent.

Conclusion

We have shown how deontological ethics can give rise to reasonably well-defined and objective principles for ethically evaluating a machine's rule base. In particular, we formulated generalization, joint autonomy, and utilitarian principles in quantified modal logic. We indicated how such a formulation can address some weaknesses of inductive logic and moral intuitions as a basis for machine ethics.

While we have not removed the human element entirely from ethical reasoning, we have removed any reliance on human moral intuitions. At the current stage of research, the application of ethical principles remains a task for the human programmer, because they are based partly on the programmer's factual knowledge base and what is rational for the programmer to believe. Humans must also identify the maximal action plans that are submitted to ethical scrutiny. Nonetheless, we have shown how deontological principles can objectively evaluate machine instructions, while calling on human judgment only in matters of fact.

Furthermore, precise moral principles are a first step toward the automation of ethical reasoning. A second step would be to implement an "ethical blocks world," in which all agents have precisely defined action plans and ethical assessment becomes a purely computational problem. This may point the way to the automation of ethics in more complex domains.

An additional advantage of a deontological approach is that it installs *reason-responsiveness*, and therefore *transparency*, in the machine. The machine can explain why it took a certain action simply by citing the maximal action plan that generates the particular rule that prompted the action. The importance of machine transparency has been much discussed in the AI literature (e.g., Fischer and Ravizza 1998; Coates and Swenson 2013; Castelvecchi 2016; Mueller 2016; Wortham, Theodorou, and Bryson 2016b; Wortham, Theodorou, and Bryson 2016b). The identification of maximal action plans is a systematic approach to developing it.

References

Alexander, J. 2012. *Experimental Philosophy*. Cambridge: Polity Press.

Anderson, M., and Anderson, S. L. 2007. Machine ethics: Creating an ethical intelligent agent. *AI Magazine* (winter) 15–26.

Anderson, M., and Anderson, S. L. 2010. Robot be good. *Scientific American* (October) 72–77.

Anderson, S. L., and Anderson, M. 2011. A prima facie duty approach to machine ethics: Machine learning of features of ethical dilemmas, prima facie duties, and decision principles through a dialogue with ethicists. In Anderson, M., and Anderson, S. L., eds., *Machine Ethics*. New York: Cambridge University Press. 476–492.

Anderson, S. L., and Anderson, M. 2014. GenEth: A general ethical dilemma analyzer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 253–261.

Anderson, M., and Anderson, S. L. 2015a. Toward ensuring ethical behavior from autonomous systems: A case-supported principle-based paradigm. *Industrial Robot: An international Journal* 42:324–331.

Anderson, S. L., and Anderson, M. 2015b. Towards a principle-based healthcare agent. In van Rysewyk, S. P., and Pontier, M., eds., *Machine medical ethics*. Springer. 67–78.

Anderson, M.; Anderson, S. L.; and Armen, C. 2006. An approach to computing ethics. *IEEE Intelligent Systems* 21:2–9.

Anderson, M.; Anderson, S. L.; and Berenz, V. 2017. A value driven agent: Instantiation of a case-supported principle-based behavior paradigm. In *Proceedings of the AAAI Workshop on AI, Ethics, and Society*, 72–80.

Anscombe, G. 1957. *Intention*. Oxford: Basil Blackwell.

Appiah, K. A. 2008. *Experiments in Ethics*. Cambridge, MA: Harvard University Press.

Beauchamp, T. J., and Childress, J. F. 1979. *Principles of Biomedical Ethics*. New York: Oxford University Press.

Bilgrami, A. 1996. *Self-Knowledge and Resentment*. Cambridg, MA: Harvard University Press.

Buchanan, A. E., and Brock, D. W. 1990. *Deciding for Others: The Ethics of Surrogate Decision Making*. New York: Cambridge University Press.

Castelvecchi, D. 2016. Can we open the black box of AI? *Nature* 538:20–23.

Coates, D. J., and Swenson, P. 2013. Reasons-responsiveness and degrees of responsibility. *Philosophical Studies* 165:629–645.

Davidson, D. 1963. Actions, reasons, and causes. *Journal of Philosophy* 60:685–700.

Fischer, J. M., and Ravizza, M. 1998. *Reasons-responsiveness and degrees of responsibility*. Cambridge: Cambridge University Press.

Guarini, M. 2011. Computational neural modeling and the philosophy of ethics: Reflections on the particularism-generalism debate. In Anderson, M., and Anderson, S. L., eds., *Machine Ethics*. New York: Cambridge University Press. 20–23.

Kant, I. 1785. *Grundlegung zur Metaphysik der Sitten (Foundations of the Metaphysics of Morals)*, volume 4 of *Königlichen Preußischen Akademie der Wissenschaften: Kants gesammelte Schriften*. Berlin: Georg Reimer (1900).

Korsgaard, C. M. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Mueller, E. T. 2016. *Transparent Computers: Designing Understandable Intelligent Systems*. CreateSpace Independent Publishing Platform.

Nagel, T. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

O'Neill, O. 2014. *Acting on Principle: An Essay on Kantian Ethics, 2nd ed.* Cambridge: Cambridge University Press.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Ross, W. D. 1930. *The Right and the Good*. Oxford: Oxford University Press.

Schwitzgebel, E., and Rust, J. 2016. The behavior of ethicists. In *A Companion to Experimental Philosophy*. Malden, MA: Wiley Blackwell.

Sinnott-Armstrong, W. 2006. Moral intuitionism meets empirical psychology. In *Metaethics after Moore*. New York: Oxford University Press.

Wallach, W.; Allen, C.; and Smit, I. 2008. Machine morality: Bottom-up and top-down approaches for modeling human moral faculties. *AI & Society* 565–582.

Wortham, R. H.; Theodorou, A.; and Bryson, J. J. 2016a. Robot transparency, trust and utility. In *Principles of Robotics Workshop*. Sheffield, UK: Proceedings of AISB.

Wortham, R. H.; Theodorou, A.; and Bryson, J. J. 2016b. What does the robot think? Transparency as a fundamental design requirement for intelligent systems. In *Ethics for Artificial Intelligence Workshop*. New York: IJCAI.