A rough proposal version: Do not cite without permission.

Toward A Non-Intuition-Based Machine Ethics

John Hooker & Tae Wan Kim
Carnegie Mellon University
June 2017

1 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, & Smit, 2005), Anderson and Anderson (hereafter, A&A) show that machine ethics is better to be principle-based (cf., Guarini, 2011). Relatedly, they also show that any adequate machine ethics should have non-consequentialist elements to minimally respect the dignity of persons. We build upon their work, but critically. A&A 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) the approach relies on human moral intuition too much; ii) the approach arbitrarily construes autonomy. We propose a non-intuition- and principled-autonomy-based machine ethics.

2 Anderson and Anderson's Prima Facie Duties Approach

A&A use an inductive logic programming to discover a decision principle for a specific domain. In specific, 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?"

There are only two options that the system can take: “Notify” or “Don’t.” Relying upon the biomedical ethicists Buchanan and Brock’s work (1990), A&A assume that the correct answer is “Do not notify.” To analyze the rationale, drawing upon the work of biomedical ethicists Beauchamp and Childress (1979), A&A assume that three *prima facie* duties—non-maleficence, beneficence, and autonomy—are relevant to the scenario. A&A use the value of each duty as an ordered triple as follows:
“Notify” (minimize harm, maximize beneficence, maximize respect for autonomy)
“Don’t” (minimize harm, maximize beneficence, maximize respect for autonomy)

In both cases—“Notify” and “Don’t”—harm is not involved. In the case of

> “Notify,” some value of beneficence is earned, while the value is lost in “Don’t.” In “Notify,” some value of autonomy is lost, whereas in “Don’t”

some value of autonomy is earned. One way to show why the right choice is “Don’t notify” is to understand that the value of respecting autonomy is +2 while its violation is -2, whereas the value of beneficence is +1 while its violation is -1, as follows:

“Notify” (0, 1, -2)

“Don’t” (0, -1, 2)

Repeating the same process for other variations that cover all the possible cases, the system inductively seeks equilibrium that coherently covers all the choices that A&A assume as correct, drawing upon Buchanan and Brock (1990). As a result, the decision principle that A&A’s system discovers in the healthcare context above 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, following the intuitionist W.D. Ross (1930), A&A’s system relies upon ethicists’ moral intuition. Ross argued that ethics consists of various duties that sometimes conflict with each other and that when duties are conflict with each other, we should rely on well-educated people’s moral intuitions. Ross (1930: 41) 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."

Consistently, 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).

Above, assuming the correctness of the ethicist's intuition—that "Do not 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.

3 Problems of Moral Intuition in Machine Ethics

But 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 (e.g., moral intuitions 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 ethicists' intuitions are not significantly different from ordinary people's (For reviews, see Schwitzgebel & Rust, 2016).[1]

Additionally, a major rationale behind developing machine ethics is inconsistent with reliance on human intuition. In analogy, a major rationale

[1] The cited paper does not survey works that directly study ethicists' moral intuition, but the consistency between their ethical beliefs and behaviors. If it is plausible to believe that humans often use intuitions to guide behavior (Haidt, 2001), the research implies that ethicists' moral intuitions are not more reliable than ordinary people's.

for developing autonomous vehicles is that a primary cause of car accidents is human mistakes, and autonomous vehicles minimize human involvement. Likewise, if human moral intuitions are not reliable, machine ethics should be developed in a way to best 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 argued that machines should not be allowed to make a choice for such cases. But machines may face scenarios in which they cannot but make choices (not making a choice is itself a choice). And, as A&A often emphasize, to make a breakthrough in dilemma cases that human ethicists cannot solve should be a contribution of machine ethics. Below, we will develop a non-intuition-based approach.

4 Individual 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 (i.e., maximize expected net pleasure). But soon they turned to the *prima facie* duty approach because of standard problems in utilitarianism: utilitarianism demands sacrificing one for the sake of the good of many, so individual persons' minimum respect is not guaranteed. We agree with A&A that machine ethics must possess deontological elements that protect a person's dignity.[2] In A&A's system, a duty to maximize respect of autonomy plays such a role. But we believe that how A&A computerize autonomy fails to fulfill its expected role.

Return to the scenario above. The value of "Notify" is (no harm: 0, maximize benefit: 1, respect autonomy: -2; net value= -1), while that of "Don't" is (0, - 1, 2; net value = 1). Here, the duty of autonomy seems to well function in its role, because there is only one patient and the patient is the only beneficiary. But imagine a case in which disrespecting one person's autonomy maximizes the benefits of tens of thousands of involved parties (e.g., lying to a client to maximize stockholder value). For the sake of argument, suppose that a (business) ethicist's intuition is that lying to the client to maximize profit is unethical. A&A's system, then, should say that the

[2] This does not mean that utilitarianism cannot be part of machine ethics. For instance, a deontic version of utilitarianism can be coherently combined with deontological principles.

value of autonomy violation in this case (and relevantly similar cases) is, say, $\leq$ -10,001, if the system thinks that the value of beneficence is 10,000. Now imagine that there is only one stockholder. Then the system should say that the value of autonomy violation is, say, -2, because the value of beneficence is only 1. We do not see a non-arbitrary reason to think that the value of the client's autonomy dramatically differs depending upon the number of parties that can be benefited by disrespecting her autonomy. This problem occurs because there is no principled notion of autonomy in A&A's system.

A&A's notion of autonomy, which they derive from Beauchamp and Childress (1979), identifies autonomy as self-control, self-governance, or independence. But such an individualist notion of autonomy is vulnerable to the question of to what extent one's autonomy must be respected. In A&A's system, an ethicist's moral intuition functions as a side constraint to determine how far one's autonomy must be respected, but as we discussed above human moral intuition is unreliable. Below, we will develop a principled autonomy approach.

5 A principled autonomy approach

The goal of machine ethics is, to use Moore's (2006) term, to develop an "explicit ethical agent," a machine that *autonomously* finds the best course of action by being able to represent the situation it faces.[3] The public worries about the *autonomous* aspect of advanced machines, but there is a sense of autonomy by which autonomous machines are inherently ethical machines (Hooker, a). We use that sense of autonomy. We are indebted to Kant, but further develop it

Although many think that it is Kant who most articulated and defended an individualist conception of autonomy, Kant never used the individualist conception. For instance, O'Neil (2002) writes:

> "He[Kant] never speaks of an *autonomous self* or *autonomous persons* or *autonomous individuals*, but rather of the *autonomy of reason*, of the *autonomy of ethics*, of the *autonomy of principles* and of the *autonomy of willing*. He does not see autonomy as something that some individuals have to a greater and others to

[3] In Moore's taxonomy (2006), humans are "full agents," so a machine ethics agent does not have to be a human-like agent in various manners (e.g., sympathy).

> a lesser degree, and he does not equate it with any distinctive form of personal independence or self-expression…" (83).

For Kant, autonomy is a *matter of acting on principles* and, specifically, principles of action that we could rationally will everyone to do. So, for Kant, an agent's act is autonomously done just when the agent acts on a principle that she could rationally will everyone to do so.

In the literature of machine ethics, Powers (2005, 2006) already attempted to computerize the principled autonomy. In what follows, we move beyond Powers's contribution in two manners. First, we add philosophical rigor by introducing a more adequate interpretation of Kant's Formula of Universal Law (FUL) to machine ethics. Second, we offer more specified ways to compute our version of FUL.

Note that our purpose is not to show that our version is more Kantian than Powers's. Our version may be a possible interpretation of Kant's own formula, but our purpose is to introduce an adequate form of FUL to machine ethics, regardless of whether it is what Kant meant. Our approach is Kantian, but not Kant's own.

6 Powers's FUL

A textbook translation of FUL is as follows:

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

Powers (2005, 2006) followed a certain interpretation of FUL, according to which it is unethical (i.e., impermissible) to act on any maxim that could not be universally willed (i.e., a maxim that is self-defeating once universalized). For instance, to use Kant's own example, consider that Jim wants to promise repayment of a loan without the intention not to repay. The maxim here is "Anyone promises (or is permitted to promise) repayment of a loan without the intention to repay." It is inconceivable to universalize the maxim, because in a world in which everyone promised in that manner, the institution of promising itself would collapse. So, according to Powers' FUL, it is unethical for Jim to promise in such a way.

But Powers's FUL faces many counter-examples (Aune, 1979: Ch. 5; Parfit, 2011: Ch. 12). Among others, a major criticism is that the FUL fails to filter out "false negatives" (Broad, 1916; Forschler, 2013). For instance, consider that Jim (who lives in the U.S.) wants to drive on the left side for safety's sake. The maxim is "Anyone drives on the left side to drive safe." It is very conceivable to imagine a world wherein everyone drives on the left side and drives safely. Thus, it is not impermissible for Jim to drive on the left side *now*, Power's FUL should say.

The problem in Powers' FUL is that it seeks contradiction within a maxim in the generalized world. This problem can be addressed if the focus of contradiction is located *between* one's acting on a maxim and the maxim being a universal law. In fact, Kant emphasized this condition by always adding "at the same time" when he expressed FUL (Kleingeld, 2017). We will introduce our version of FUL that incorporates the simultaneity condition below.

7 Introducing the Generalization Test

Assuming that a growing number of companies will use artificially intelligent machines, we define the test of principled autonomy, which we call the generalization test, as follows:

> Generalization test (GT): Business entity x (e.g., person, corporation, etc.)'s action is ethical only if x can rationally believe that its reason(s) for the action is consistent with the assumption that every business entity who had the same reason(s) in relevantly similar circumstances performed the action (if it could).[4]

According to GT, the fact that the universalized form of a maxim is internally self-defeating is not itself evidence showing that it is impermissible to act on the maxim. The point of GT is that an agent ought to be able to will a maxim and at the same time, consistently, to will the maxim as a universal law. Although in a world in which everyone drove on the left side, Jim could drive on the left side and drive safely, Jim cannot drive on the left side safely now. So he cannot consistently will both conditions at the same time.

We now show how the GT-Equipped Machine (GTEM) works without relying upon moral intuition and in a principled manner. Imagine that an

[4] For a detailed illustration and defense of this principle, see Hooker (b).

investment bank, MoneyRocks, uses an AI-based machine to design a financial investment instrument to make profit. The bank wants (or is mandated by an authority to have) the machine not to make unethical decisions. So the bank develops the machine as a GTEM. The GTEM autonomously calculates, forecasts, and designs profitable financial products. The GTEM allows a risky product only if creating such a risky product passes GT.

Imagine that the machine has recently designed a product in which "highly complicated financial instruments are interconnected in intricate, hard-to-parse ways that make the implications and risk profile of the investment difficult to discern" (Scharding, 2015: 245)—often called the abstruse investment (e.g., the synthetic collateralized debt obligation, or synthetic CDO, which was heavily traded before the financial crisis in 2008). GTEM now asks whether the product passes the generalization test as follows:

Step 1: Pre-analysis
The particular belief: "MoneyRocks invests in AFIS to make profit."
Let mr = MoneyRocks
Let Rx = x achieves target reason ("making profit")
Let Ax = x performs the given behavior ("investing in abstruse financial investments (AFIs).

$$\exists mr[Amr \rightarrow Rmr]$$

Step 2: Generalization
"Every business enterprise in similar circumstances invests in financial instruments in every opportunity in which the enterprise could do so."

$$\forall x[Ax \rightarrow Rx]$$

Step 3: Checking the consistency: The GTEM asks whether the company MoneyRocks can consistently believe both.

$$\{\exists mr[Amr \rightarrow Rmr]\} \,\&\, \{\forall x[Ax \rightarrow Rx\}$$

The GTEM consistently believes both at the same time just when the truth-value of each is true. According to Powers's FUL, the decision maker needs to immediately recognize whether it is *impossible* to conceive the truth-value of a proposition, but we believe that such an understanding faces a problem. For instance, it may be not impossible for one to promise repayment without the intention to repay even in a world in which everyone does so (for instance, humans can be deceived). But the probability of achieving the target reason must be significantly lower than that in the non-generalized world. So we suggest understanding GT in a probabilistic manner as follows:

The probability that there exists an x (MoneyRocks in our case) such that if x performs the given behavior if x can then x achieves the target reason for the behavior is isomorphic to the probability that if every x performed the given behavior if it could then x would achieve the target reason.

$$\mathrm{P}\left(\exists x[Ax \to Rx]\right) \approx \mathrm{P}\left(\forall x[Ax \to Rx]\right)$$

According to the above schema, the probability of each must be isomorphic to each other in the sense that the probability of each must be high enough in the non-generalized world and the generalized world.

Step 5: Final decision

$$Ax \to \{\mathrm{P}(\exists x[Ax \to Rx]) \approx \mathrm{P}(\forall x[Ax \to Rx])\}$$

If the isomorph condition is not met, the target action is impermissible.

How can the GTEM answer the question about the probabilities? The question involves an empirical forecast, and there are several options to consider:

i) Advanced artificial intelligence:
An advanced AI that has capabilities to forecast (specific to a particular domain), using a machine learning analysis that predicts the future based on the past financial market's history, e.g., can answer the probabilistic questions. Through classification and reclassification, the AI can find relevantly similar events in historic records or case studies.

ii) crowd sourcing/collective intelligence:
Crowd sourcing can be used, but a disadvantage is that the public does not have expertise in professional issues.

iii) the user's judgment:
Imagine a dialogue between the GTEM and a group of managers in MoneyRocks.

> *GTEM*: Can you rationally that the probability of MoneyRocks makes profits through AFIs, and at the same time, that every business entity in similar circumstances makes profit through AFIs?
>
> *Users*: Yes in the non-generalized world, but No for the generalized world (based on our experience, especially the 2008 financial crisis).
>
> *GTEM*: MoneyRocks is prohibited to invest in AFIs.

The user's judgment has the advantage that the answer is based on direct knowledge and experience, but it is susceptible to impartial judgment. As an alternative, experts' opinions can be used.

iv) reliance on third party experts' opinions:
The same kind of dialogue as above is used, except that users are third part experts. For instance, SEC can require MoneyRocks to show that the probability for the generalized world is high enough with a background simulation data about how a finance market works, provided by the SEC (by doing so, the SEC can regulate risky financial products).

One might say that these solutions all rely on human opinions (Correct, even the machine learning method must begin with data created by humans). So, the reader might say that our approach is not really human bias-free, unlike the ambition that we promised. However, our principled autonomy approach is human *moral*-intuition-free, and that is what we promised. Unlike A&A's system that uses human ethicists' moral intuition, our system does not use any moral intuition. It uses only humans' non-moral/empirical knowledge. It is the beginning of a non-intuition-based machine ethics.

References

Alexander, J. (2012). *Experimental philosophy.* Cambridge: Polity Press.

Anderson, M., & Anderson, S. L. (2014). GenEth: A general ethical dilemma analyzer. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence.* Quebec.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, Winter*, 15-26.

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

Anderson, M., & 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, M., Anderson, S. L., & Armen, C. (2006). An approach to computing ethics. *IEEE Intelligent Systems, 21*, 2-9.

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

Anderson, S. L., & 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 M. Anderson, & S. L. Anderson, *Machine ethics* (pp. 476-492). New York: Cambridge University Press.

Anderson, S. L., & Anderson, M. (2015b). Towards a principle-based healthcare agent. In S. P. van Rysewyk, & M. Pontier, *Machine medical ethics* (pp. 67-78). Switzerland: Springer.

Appiah, K. A. (2008). *Experiments in ethics.* Cambridge: Harvard University Press.

Aune, B. (1979). *Kant's theory of morals.* Princeton: Princeton University Press.

Beauchamp, T. J., & Childress, J. F. (1979). *Principles of biomedical ethics.* New York: Oxford University Press.

Broad, C. D. (1916). On the function of false hypotheses in ethics. *International Joural of Ethics*, 26, 377-97.

Buchanan, A. E., & Brock, D. W. (1990). *Deciding for others: The ethics of surrogate decision making.* New York: Cambridge University Press.

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking.* New York: W.W. Norton & Company.

Forschler, Scott. (2013). Two dogmas of Kantian ethics. *Journal of Value Inquiry*, 47, 255-269.

Guarini, M. (2011). Computational neural modeling and the philosophy of ethics: Reflections on the particularism-generalism debate. In M. Anderson, & S. L. Anderson (Eds.), *Machine ethics* (pp. 316-334). New York: Cambridge University Press.

Hill, Jr., T. E. (1992). *Dignity and practical reason in Kant's moral philosophy.* Ithaca: Cornell University Press.

Hooker, J. N. (a) Autonomous machines are the best kind, because they are ethical. *Working paper.*

Hooker, J. N. (b) *Ethical decisions: Doing ethics with our brains. Working book manuscript.*

Kleingeld, P (2017). Contradiction and Kant's formula of universal law. *Kant-Studien*, 108, 89-115.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems,* 21, 18-21.

O'Neil, O. (2002). *Autonomy and trust in bioethics.* New York: Cambridge University Press.

Parfit, D. (2011). *On what matters* (Volume One) New York: Oxford University Press.

Powers, T. M. (2005). Deontological machine ethics. In S. L. Anderson, & C. Armen (Eds), *AAAI Fall Symposium Technical Report.*

Powers, T. M. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems* 21, 46-51.

Rawls, J. (1971). *A theory of justice.* Cambridge: Harvard University Press.

Ross, W. D. (1930). *The right and the good.* Oxford: Oxford University Press.

Scharding, T. K. (2015). Imprudence and immorality: A Kantian approach to the ethics of financial risk. *Business Ethics Quarterly*, 25, 243-265.

Schwitzgebel, E., & Rust, J. (2016). The behavior of ethicists. In J. Sytsma, & W. Buckwalter, *A companion to experimental philosophy* (pp. 225-233). Malden, MA: Wiley Blackwell.

Sinnott-Armstrong, W. (2006). Moral intuitionism meets empirical psychology. In T. Horgan, &M. Timmons (Eds.), *Metaethics after Moore* (pp. 339-366). New York: Oxford University Press.

Wallach, W., Allen, C., & Smit, I. (2005). Machine morality: Bottom-up and top-down approaches for modeling human moral faculties. In M. Anderson, S. L. Anderson, & C. Armen, *Machine ethics* (pp. 94-102). AAAI Press.