Truly Autonomous Machines Are Ethical

John Hooker, Tae Wan Kim

■ *There is widespread concern that as machines move toward greater autonomy, they may become a law unto themselves and turn against us. Yet the threat lies more in how we conceive of an autonomous machine rather than the machine itself. We tend to see an autonomous agent as one that sets its own agenda, free from external constraints, including ethical constraints. A deeper and more adequate understanding of autonomy has evolved in the philosophical literature, specifically in deontological ethics. It teaches that ethics is an internal, not an external, constraint on autonomy, and that a truly autonomous agent must be ethical. It tells us how we can protect ourselves from smart machines by making sure they are truly autonomous rather than simply beyond human control.*

As companies and governments race to develop autonomous systems, such as self-driving vehicles, robotic caregivers, and autonomous weapons, we worry about losing control of our machines (Vinge 1993; Bostrom 2014; Smith and Anderson 2017). We imagine an autonomous agent to be one that makes its own decisions, free of external constraints, including ethical constraints. Consequently, we fear that autonomous machines will become oblivious to our interests and welfare. As artificial intelligence (AI) systems become increasingly intelligent, and increasingly embedded in almost every aspect of our lives, the worry intensifies.

 ISSN 0738-4602

Yet there is a sense of autonomy, deeply rooted in the ethics literature, according to which an autonomous machine cannot be unethical. In this tradition, ethics imposes internal (as opposed to external) constraints on autonomous action, because ethical obligation is bound up in the very concept of autonomy (Nagel 1986; Korsgaard 1996; Bilgrami 2006; O'Neill 2014). This idea derives from a thought tradition in ethics known as deontology. Although more than two centuries old, it is remarkably well equipped to deal with the coming age of superintelligent machines, because it grounds ethics in the logical structure of action without presupposing that the agent is human.

Autonomy Versus Independence

Etymologically, autonomy means *self-law*, which may give the impression that an autonomous agent is a law unto itself, free of constraints. This is reflected in the most widely cited definitions in the AI literature (for a survey, see Beer, Fisk, and Rogers 2012). One goes as follows:

> An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of *its own agenda* and so as to effect what it senses in the future. (Franklin and Graesser 1996, p. 25, emphasis added.)

Another definition strikes a similar tone:

> Autonomous agents possess goals which are *generated* from within rather than *adopted* from other agents. These goals are generated from *motivations* which are higher-level non-derivative components characterizing the nature of the agent, but which are related to goals. (Luck and d'Inverno 1995, p. 258, original emphasis)

> A motivation is *any desire or preference* that can lead to the generation and adoption of goals and that affects the outcome of the reasoning or behavioural task intended to satisfy those goals. An autonomous agent is an agent with a non-empty set of motivations. (Luck and d'Inverno 2001, p. 13, emphasis added.)

Hui-Min Huang et al. (2007) take basically the same view in the Autonomy Levels for Unmanned Systems workshop series, where they treat the level of autonomy as the level of human/operator independence.

Autonomous machines conceived in this fashion are indeed a threat as well as an opportunity because they choose their own goals independently of constraints. Even if they do not choose ultimate goals, they find their own means to achieving goals they are assigned. A caregiver robot can potentially put its owner's cat into the microwave to prepare dinner.

These definitions fail to recognize that autonomy must incorporate an element of rationality. Suppose that you ask a young student, "What is 2 + 2?" The child proudly answers, "5," boasting that this is his autonomous choice. You try to explain that rationality requires us to accept the answer "4" by pointing out that two baskets of two oranges contain four oranges. Yet the student keeps refusing to compromise his autonomy.

Now suppose that you ask a student, "Is it ethical for you to harm the innocent?" She answers, "Absolutely, yes," boasting that that is her autonomous choice. In both cases, the student is confused about what autonomy means. Kantian philosopher Alan Donagan (1984) once aptly pointed out, "The notion that an autonomous being is one having the power to do as it likes is a vulgarity" (p. 129). Autonomous choices must have a rational basis, and the task of deontological ethics is to understand what that basis is.

The AI literature is not entirely indifferent to the importance of rationality. Wooldridge in his widely cited book on multiagent systems remarks,

> Of course, we do not have complete freedom over beliefs, goals, and actions. For example, I do not [rationally] believe I could choose to believe that 2 + 2 = 5; nor could I choose to want to suffer pain. Our genetic makeup, our upbringing, and indeed society itself have effectively conditioned us to restrict our possible choices (Wooldridge 2009).

However, Wooldridge does not further develop this aspect of autonomy and, in the end, endorses the dominant view of autonomy as independence. Russell and Norvig's influential textbook on AI states, "A rational agent is one that does the right thing. A rational agent should be autonomous — it should learn what it can to compensate for partial or incorrect prior knowledge" (Russell and Norvig 2003, pp. 36, 39). They clearly imply that rationality is not determined by subjective preferences when they say, "If we define success in terms of agent's opinion of its own performance, an agent could achieve perfect rationality simply by deluding itself that its performance was perfect" (p. 37). However, the connection between rationality and autonomy is not discussed elsewhere in the textbook, which settles on the dominant notion of autonomy as independence.

What Is Autonomous Action?

Our core argument can be sketched in the following steps: that an action is autonomous if the agent's reasons for the action can explain the action; that an agent's reasons for an action can explain the action only if they are coherent; that ethical principles are nothing more than necessary conditions for the coherence of the reasons; and that, therefore, autonomous actions cannot be unethical.

We begin by developing the concept of autonomous action used in our core argument. We define autonomous action to be behavior that, at least potentially, has two kinds of explanation. On the one hand, it can be explained as the result of a biologic mechanism, or electronic circuitry that implements an algorithm or a multilayer neural network. On the

other hand, it can also be reasonably explained as the outcome of a process of deliberation in which the agent adduces reasons for the behavior. A piece of behavior that has this kind of dual explanation is an action.

In this sense, an insect does not act. If a mosquito bites me, its behavior can be explained only as the result of chemistry and biology. It is unreasonable to suppose that the mosquito thought to itself, "I am really hungry for blood tonight, I can satisfy my hunger by injecting my proboscis into that human's body, and I will therefore buzz over and do so."

Human behavior may also fail to be action. My hiccup is not an act because, while it has gastric causes, one cannot reasonably say that I chose to hiccup for some particular reason. Nonetheless I am an agent because I am capable of action. If I hold my breath in an attempt to stop the hiccups, there are presumably complex neurologic causes for my behavior, but it can also be explained as the result of ratiocination. Perhaps I reasoned that because I have often been told that holding one's breath can stop hiccups, there may be some truth to this, and because hiccups are annoying, I may as well give it a try. My reasons need not be good or convincing reasons, but it must be reasonable to attribute them to me, and they must be coherent enough to count as an explanation for why I held my breath. This is the sense in which autonomous action must have a rational basis.

Actions and Reasons

The connection between action and having reasons is deeply embedded in the philosophical tradition, having origins in the work of Immanuel Kant and perhaps ultimately in Aristotle. In recent decades, this connection has become part of what might be regarded as a textbook account of agency, beginning with Anscombe (1957) and Davidson (1963).[1]

This account poses a problem, however. Actions resulting from a reasoning process are every bit as determined as other behavior, because the reasoning process is itself determined. This raises the ancient conundrum of freedom versus determinism, which recent neurologic experiments have revived. An MRI machine can detect changes in the brain that take place a few seconds before one's decision to take an action, such as moving a finger (Soon et al. 2008). We may have the impression of making a decision, but this is false consciousness. Brain chemistry and its causal antecedents have already made the decision for us. It may appear, then, that there can be no free choice, and therefore no ethics.

In a previous issue of *AI Magazine*, Covrigaru and Lindsay (1991) pointed out that once we accept the standard scientific view (i.e., determinism), autonomy as *independence* — the dominant understanding of autonomy in the AI literature — becomes "illusive." To overcome this problem, they proposed a notion of "deterministic autonomous systems," according to which autonomy is defined not as a real property but merely as a subjectively perceived state that humans attribute to a system that exhibits certain kinds of goal-directed behavior. Yet if autonomy is illusory, it remains unclear how free choice and ethics are possible.

It is to overcome this quandary that autonomous action is conceived as having two kinds of explanation — causal and reasons-based. We can view ourselves as acting autonomously, even while physically determined, if we identify reasons behinds our actions that can be evaluated from an ethical perspective. This idea, too, has roots in Kant. As he put it, "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*" (Kant 1785, original emphasis). In other words, to see ourselves as taking action (in Kantian language, to think of ourselves as "practical"), we must interpret ourselves as existing in a "world of understanding" outside the natural realm of cause and effect. Or to use more modern language, we must be able to give our behavior a second kind of explanation, one that is based on reasons we adduce for it rather than cause and effect. This idea eventually evolved into the "dual standpoint" theories of recent decades (Nagel 1986; Korsgaard 1996; Bilgrami 2006).

A dual standpoint theory may not fully resolve the problem of freedom versus determinism (Nelkin 2000), but it sets the stage for ethics. It provides a well-defined criterion for distinguishing autonomous action from mere behavior, and for distinguishing agents from nonagents, and this is all we need. We cannot offer a full-blown argument for a dual standpoint theory here, but we invite readers to see what happens once we accept the premise. The proof of the pudding is in the eating.

From Action Theory to Ethics

The theory of action just sketched leads immediately to ethical principles. Recall that an agent's reasons for an action can explain the action only if they are coherent. Ethical principles are nothing more than necessary conditions for the coherence of the reasons.

While a number of necessary conditions for logical coherence might be stated, those most relevant to ethics are derived by appealing to the universality of reason: The validity of one's reasons 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.

To see how this premise can lead to an ethical principle, suppose that I tell lies simply because it is convenient to deceive people. 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 agents, or

as Kant would say, I must regard my choice of action as "legislating" a general policy for everyone. This is captured in the famous generalization principle, which is perhaps best stated as follows:

Generalization Principle

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.

Onora O'Neill (2014) provides an excellent reconstruction of the Kantian argument for this principle.[2] Suppose again that I tell lies because it is convenient to deceive people, which means that 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. This does not mean that others would in fact lie for mere convenience if I decide to do so. It only means that my reasons for lying are inconsistent with the assumption that others lie for the same reasons.

In other words, the rational process behind my decision to lie is self-contradictory. I am adopting a policy of lying when deception is convenient, but at the same time I am not adopting a policy of lying when deception is convenient, because adopting this policy means adopting it for everyone, which I am rationally constrained to believe defeats my purpose in lying. Because of this logical contradiction, my reasons cannot be taken as an explanation for my behavior. They need not be good reasons or convincing reasons, but they must be coherent enough for one to see them as explaining why I did what I did.[3]

The generalization principle is more sophisticated and nuanced than it may first appear. It can condone lying, for instance, under the right circumstances. To take a famous case, employees in an Amsterdam office building lied to the Nazi state police when asked the whereabouts of Anne Frank and her family. They told the police that they had no idea, even though the family was holed-up in that very building. The reason for the lie was that it would avoid tipping off the police. This purpose would still be achieved if everyone with this reason to lie did so, even if the police did not believe the lies. So, the lie conforms to the generalization principle. The principle also tells us when we should not break a promise, breach a contract, steal, release online data to marketing firms, and so forth with countless other actions.

Humans and machines are equally bound by the generalization principle because its derivation is based only on the formal properties of agency, not on whether the agent is human. Thus, autonomous machines will not deceive us or break agreements with us, for example, in circumstances in which we humans are obligated not to lie or break agreements.

Similar lines of thought lead to additional ethical principles, such as respecting the autonomy of other agents. The key point here is that violation of these principles means that the agent's reasoning is incoherent. It is impossible to explain the agent's behavior as based on reasons, and therefore to regard it as action. All actions are ethical if they are truly autonomous actions and not mere behavior. The ethical imperative is, in essence, a call to exercise one's capacity for autonomous action. This is why a truly autonomous machine is an ethical machine.

Respecting the Autonomy of Others

A second ethical principle that binds both humans and machines is respect for the autonomy of other agents. This principle ensures that autonomous machines will not take over or otherwise oppress us in unethical ways.

The argument for respecting autonomy, in a nutshell, is this. Suppose I violate someone's autonomy for such-and-such reasons. That person could, at least conceivably, have the same reasons to violate my autonomy. This means that, due to the universality of reason, I am endorsing the violation of my own autonomy in such a case. This is a logical contradiction, because it implies that I am deciding not to do what I decide to do. My violation of another agent's autonomy therefore makes the reasoning behind my behavior incoherent, and so it cannot be viewed as autonomous action.

Respecting the autonomy of other agents does not mean allowing them to do anything they want. To understand this, we must take a few moments to develop the principle more carefully.

First, we note that decisions to act have a conditional character. Because these decisions are based on reasons, they are decisions to act if the reasons apply. For example, if you decide to cross the street to catch a bus at the bus stop, your decision perhaps has the form, "If I want to catch a bus, and the bus stop is across the street, and no cars are coming, then I will cross the street." Let's call this sort of conditional decision an action plan. Machine conduct can be properly evaluated only if it is programmed as action plans, but this is actually convenient, because the conditional form is quite natural for machine instructions.

We can now see why respect for autonomy can permit a certain amount of coercion. Suppose you begin to cross the street toward the bus stop, unaware that a car is approaching. Your robot companion shouts a warning, and when you do not hear, it forcibly pulls you out of the path of the car. This is not a violation of your autonomy, because it is consistent with your action plan of crossing the street if no car is coming. This is recognized by the principle of respecting autonomy:

Principle of Respecting Autonomy

It is unethical for me to select an action plan that I am rationally constrained to believe interferes with an action plan of another agent.

The concept of action plan also allows interference when there is informed consent because consent in effect becomes part of the action plan. Suppose a robot performs surgery on me that leads to complications, thwarting my plan to travel next month. However, I signed a release that permits surgery, knowing that complications could result. This modified my action plan for travel, which became, "If there are no complications from surgery, then travel next month." The robotic surgeon therefore did not interfere with my action plan. So, we have the principle of informed consent:

Principle of Informed Consent

Interfering with an agent's action plan is no violation of autonomy when that agent has given informed consent to the possibility of interference, and giving this consent is, itself, a coherent action plan.

Finally, the obligation to respect autonomy does not forbid interfering with unethical behavior, in the sense of behavior that violates other ethical principles, because unethical behavior is not an exercise of agency in the first place. This leads to a companion principle, the interference principle:

Interference Principle

Coercion that prevents only unethical behavior does not compromise autonomy.

If your robot companion grabs your arm when you attempt to steal someone's smart phone, there is no violation of your autonomy, because theft is (normally) ungeneralizable and therefore unethical. However, if your robot locks you in a closet to prevent you from writing false numbers on your income tax form, it violates your autonomy, even though income tax evasion is unethical. Being locked in a closet prevents you from performing any number of ethical actions. A more extensive analysis of when restraint is justified, based on a concept of joint autonomy, can be found in Hooker (2018).

Building Ethical Machines

Nothing in this essay is meant to imply that autonomous machines can or should be developed. It only argues that if we move in this direction, we can make sure the machines are ethical by making them truly autonomous, as opposed to merely independent of human control. In the meantime, deontological analysis can be a valuable guide to building machines that are ethical but not yet autonomous (Hooker and Kim 2018). We need only apply ethical principles to the human designer of the machine rather than the machine itself.

It is useful to think about how to design an ethical machine, because this will help us understand how a truly autonomous machine must be structured. First, for us to apply ethical principles, the machine must be ultimately governed by action plans; that is, by if-then rules that instruct the machine to perform certain actions in certain circumstances. The antecedent (if-part) of a rule is interpreted as the reason for the action, and the ethical tests applied on that basis.

If we are to apply the tests properly, the antecedent must capture the true reason for the action, in full generality. Suppose, for example, that a self-driving ambulance is instructed to use sirens and lights in a medical emergency. This is acceptable, but the designer has inserted additional instructions of the form:

If a patient needs nonemergency transport from location *X* to location *Y* between 9 and 10 AM, and if using siren and lights would result in faster delivery, give the patient a ride with siren and lights using route *Z*.

There are instructions for each pair of locations and each time of day because different routes are optimal in each case. Nonemergency use of siren and lights (to save time) violates the generalization principle, because if it were generalized, other drivers would simply ignore ambulances, and the siren and lights would not save time. Yet each of the instructions is generalizable because the conditions are so specific that they apply only occasionally and would have no effect on the behavior of other drivers. The problem is that the scope of the antecedent is too narrow. The real reason the sirens and lights are to be used is to transport patients more rapidly. The specific instructions are derived from the general action plan

If a patient needs nonemergency transport, and if using siren and lights would result in faster delivery, then give the patient a ride using siren and lights along the optimal route.

This is the action plan that must be subjected to ethical scrutiny, and it is not generalizable.

AI systems frequently use multilayer neural networks to select actions. This might be captured in an action plan like

If a neural network of a certain architecture, trained in a certain way on a certain data set, indicates that a patient should be transported using siren and lights, then transport the patient using siren and lights.

To check this action for generalizability, the designer must investigate the results of operating all ambulances as dictated by the neural network. This could be difficult to assess, due to the nontransparency of the network. Nonetheless questions of this sort must be answered if deep learning is to provide an ethical basis for machine behavior.

Instructions must also be evaluated with respect to whether they violate the autonomy of other agents. To adapt an example from Anderson and Anderson (2007, 2011), suppose a robotic assistant in a nursing home administers medication to patients. It is given the instruction.

> If you offer prescribed medication *X* to patient *Y* at the appropriate time, and the patient refuses to take it, then inform the nursing staff.

The patient insists that she has the right to control what goes into her body and does not wish the nursing staff to be informed of her refusal. Ignoring her wishes may seem to be a violation of autonomy, but it is not, because it neither compels her to take the medication nor interferes with any other ethical action plans. Her desire to keep her refusal secret is not an action plan, ethical or otherwise. It is only a desire, and the autonomy principle does not require us to grant a wish simply because someone desires it.

Now suppose that medication *X* is necessary to prevent the patient *Y* from becoming disoriented. The nursing staff confines disoriented patients to the building, because otherwise they may suffer an accident on the busy streets outside. If the patient plans to leave the building while disoriented — perhaps she has a coherent reason for taking the risk — the aforementioned instruction violates the autonomy principle. A modified instruction, however, could pass muster due to the principle of informed consent:

> If you offer prescribed medication *X* to patient *Y* at the appropriate time, the patient refuses to take it, and the patient autonomously gave informed consent to a policy of informing the nursing staff of such refusals when she voluntarily entered the nursing home, then inform the nursing staff.

Further refinements of the instruction may be necessary in a realistic setting, but we at least have a fairly precise guide for evaluating its ethical status.

Building an Autonomous Machine

A truly autonomous machine formulates action plans as well as following them. To create an action plan, the machine must supply the reasons that comprise the antecedent of the action plan, and those reasons must be coherent enough to explain why the resulting action is undertaken. In particular, they must satisfy the generalization principle and respect autonomy.

Transparency and explainability are therefore essential characteristics of an autonomous machine. If a machine's every action must result from an action plan, then the machine must be reasons-responsive. It must be able to provide a coherent reason for every action to formulate the action plan. The practical importance of transparency and explainability in AI has been much discussed (Mueller 2016; Wortham, Theodorou, and Bryson 2016a,b). We now see that it is not only important but bound up in the very concept of an autonomous agent.

We can also begin to see what kinds of abilities are required for genuine autonomy. If an autonomous AI system is to rely on deep learning and neural networks, for example, these networks must deliver not only action choices but reasons for the actions. Furthermore, the system must be able to determine whether the resulting action plans (or more precisely, the overarching plans from which the more specific plans derive) satisfy the generalization and other principles. This requires that the system carry out thought experiments, which in turn rely on its ability to accumulate beliefs about matters of fact and assess whether they are rational. For example, a truly autonomous ambulance must be able to determine whether it is rationally constrained to believe that drivers would ignore ambulances if they all abused the siren and lights.

None of this implies that truly autonomous machines must acquire such human traits as feelings, sympathy, loyalty, or intellectual curiosity. They need only exhibit the formal properties of agency. Yet, as we see, building these properties into a machine is an extremely daunting challenge. The challenge may eventually be met, but perhaps only in such limited domains as driving, household chores, or certain personal services.

Implications for Policy and Standards

Current laws define autonomous systems in terms of independence. For instance, California Senate Bill 1298 (Chapter 570), which authorized the Department of Motor Vehicles to develop regulations for the testing and operation of autonomous vehicles, defines "autonomous technology" as "technology that has the capability to drive a vehicle without the active physical control or monitoring of a human operator" and "autonomous vehicle" as "any vehicle with autonomous technology" (Division 16.6). Autonomous vehicles defined in this manner can indeed present a threat to humans, as discussed at the beginning of this essay.

Most major standards for the safety and ethics of AI likewise equate autonomy with independence. For instance, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems endorses the definition of an autonomous weapon system offered by the International Committee of the Red Cross:

> … a system that can select (that is, search for or detect, identify, track, select) and attack (that is, use force against, neutralize, damage or destroy) targets without human intervention (IEEE 2018, p. 116).

To guard against marauding machines, AI policies and standards should take account of true autonomy as well as independence, and make sure that one accompanies the other. Laws can mandate that a code of ethics be programmed into machines, but to the extent that the machines are independent, they can ignore such admonitions.

The tension between autonomy (as popularly conceived) and ethics can be resolved only through a unified approach that recognizes the fundamental connection between the two.

Concluding Remarks

We know instinctively that we must be very careful about endowing machines with the power of choice. Deontological ethics tells us in precisely what sense we must be careful: We must ensure that our increasingly intelligent machines have the capacity for true autonomy as well as independence.

One might object that we have offered no real, engineering solutions to the threats potentially posed by autonomous machines. Such an objection misunderstands our thesis. Our position is that a first step toward a real solution is not more sophisticated engineering, but a more sophisticated concept of autonomy. We need a revolution in thought.

Notes

1. Contemporary philosophers Pettit and Smith (1996) advance a related thesis that autonomy is inherently responsive to reason. To contrast mere independence from reason-responsive autonomy, they dub the latter "orthonomy." Our reconceptualization of Kantian autonomy is somewhat similar to their orthonomy, but a detailed comparison is beyond the scope of this essay.

2. The generalization principle is closely related to Kant's Categorical Imperative, which is notoriously subject to interpretation. The Imperative appears in our development as the universality of reason, of which the generalization principle is seen as a direct consequence. An interpretation of the Imperative that is not based on the reasons for action is LN4 of Parfit (2011, p. 317).

3. While behavior must be explicable as based on reasons to qualify as action, this does not mean that irrational factors like emotion or feelings can play no role in the rationale for an action. Suppose I avoid driving over a certain bridge because I had a serious accident there at some point in the past. My avoidance of the bridge is an action if I can explain, in some coherent fashion, why I avoid the bridge. Perhaps the memory of my accident makes me feel nervous to drive over the bridge, and it is unpleasant to feel nervous. My aversion to driving over the bridge may be irrational in some sense, particularly if the bridge is as safe as any other route. Yet my rationale is a coherent explanation for my avoidance. It may be ethical as well, unless (for instance) a refusal to use the bridge prevents me from carrying out obligations on the other side. On the other hand, if I simply avoid the bridge without adducing any reasons why — reasons that can be checked for coherence — then my avoidance is not an action. In this case, my unpleasant memory of the accident is merely a cause for my avoidance rather than a reason for it. Similarly, the output of a robot's neural network is a cause of the robot's behavior, rather than a reason for it. To be autonomous, the robot must generate reasons for its behavior that can be put to the test ethically.

References

Anderson, M., and Anderson, S. L. 2007. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* 28(4): 15–26.

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 *Machine Ethics*. M. Anderson, and S. L. Anderson, editors. 476–92. New York: Cambridge University Press. doi.org/10.1017/CBO9780511978036.032.

Anscombe, G. E. M. 1957. *Intention*. Oxford, UK: Basil Blackwell.

Beer, J. M.; Fisk, A. D.; and Rogers, W. A. 2012. *Toward a Psychological Framework for Levels of Robot Autonomy in Human-Robot Interaction. Technical Report*. Atlanta, GA: Georgia Institute of Technology.

Bilgrami, A. 2006. *Self-Knowledge and Resentment*. Cambridge, MA: Harvard University Press.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press.

Covrigaru, A. A., and Lindsay, R. K. 1991. Deterministic Autonomous Systems. *AI Magazine* 12(3): 110.

Davidson, D. 1963. Actions, Reasons, and Causes. *The Journal of Philosophy* 60(23): 685–700. doi.org/10.2307/2023177.

Donagan, A. 1984. *Justifying Legal Practice in the Adversary System*. Lanham, MD: Rowman and Allanheld.

Franklin, S., and Graesser, A. 1996. Is It an Agent, or Just a Program? A Taxonomy for Autonomous Agents. In *International Workshop on Agent Theories, Architectures, and Languages*, 21–35. Berlin: Springer.

Hooker, J. N. 2018. *Taking Ethics Seriously: Why Ethics Is an Essential Tool for the Modern Workplace*. Abingdon, UK: Taylor & Francis. doi.org/10.4324/9781315097961.

Hooker, J. N., and Kim, T.-W. 2018. Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic. In Proceedings of the First *Association for the Advancement of Artificial Intelligence (AAAI)/Association for Computing Machinery (ACM) Conference on Artificial Intelligence, Ethics and Society (AIES)*. New York: Association for Computing Machinery. doi.org/10.1145/3278721.3278753

Huang, H.-M.; Pavek, K.; Ragon, M.; Jones, J.; Messina, E.; and Albus, J. 2007. Characterizing Unmanned System Autonomy: Contextual Autonomous Capability and Level of Autonomy Analyses. In *Unmanned Systems Technology IX*. Vol. 6561, 65611N. Bellingham, WA: International Society for Optics and Photonics.

IEEE. 2018. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Kant, I. 1785. *Groundwork of the Metaphysics of Morals*. Akademie edition. Vol. 4, 458. Berlin: Walter de Gruyter.

Korsgaard, C. M. 1996. *The Sources of Normativity*. Cambridge, UK: Cambridge University Press. doi.org/10.1017/CBO9780511554476.

Luck, M., and d'Inverno, M. 1995. A Formal Framework for Agency and Autonomy. In *Proceedings of the First International Conference on Multiagent Systems*. 254–60. Cambridge, MA: The MIT Press.

Luck, M., and d'Inverno, M. 2001. A Conceptual Framework for Agent Definition and Development. *The Computer Journal* 44(1): 1–20. doi.org/10.1093/comjnl/44.1.1.

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

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

Nelkin, D. K. 2000. Two Standpoints and the Belief in Freedom. *The Journal of Philosophy* 97(10): 564–76. doi.org/10.2307/2678468.

O'Neill, O. 2014. *Acting on Principle*: *An Essay on Kantian Ethics*. Second edition. Cambridge, UK: Cambridge University Press.

Parfit, D. 2011. *On What Matters*. Vol. 1. Oxford, UK: Oxford University Press.

Pettit, P., and Smith, M. 1996. Freedom in Belief and Desire. *The Journal of Philosophy* 93(9): 429–49. doi.org/10.2307/2940892.

Russell, S. J., and Norvig, P. 2003. *Artificial Intelligence*: *A Modern Approach*. Toronto: Pearson.

Smith, A., and Anderson, M. 2017. *Automation in Everyday Life*. *Technical Report*. Washington, DC: Pew Research Center.

Soon, S.; Brass, M.; Heinze, H.-J.; and Haynes, J.-D. 2008. Unconscious Determinants of Free Decisions in the Human Brain. *Nature Neuroscience* 11: 543–5. doi.org/10.1038/nn.2112.

Vinge, V. 1993. The Coming Technological Singularity: How to Survive in the Post-Human Era. Paper presented at Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace. NASA Lewis Research Center, Cleveland, OH, March 30–31, 1993. ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022856.pdf.

Wooldridge, M. 2009. *An Introduction to Multiagent Systems*. New York: John Wiley & Sons.

Wortham, R. H.; Theodorou, A.; and Bryson, J. J. 2016a. Robot Transparency, Trust and Utility. Paper presented at The Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB) Workshop on Principles of Robotics, Sheffield, United Kingdom, April 4. researchportal.bath.ac.uk/en/publications/robot-transparency-trust-and-utility.

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 *Proceedings of the International Joint Conference on Artificial Intelligence* (*IJCAI*): *Workshop on Ethics for Artificial Intelligence*. Bath, UK: University of Bath.

John Hooker is the T. Jerome Holleran Professor of Business Ethics and Social Responsibility and a professor of operations research at Carnegie Mellon University.

Tae Wan Kim is an associate professor of ethics at the Tepper School of Business, Carnegie Mellon University.