Journal of Artificial Intelligence Research 70 (2021) 871-890 Submitted 11/2020; published 02/2021

Taking Principles Seriously: A Hybrid Approach to Value Alignment in Artificial Intelligence

Tae Wan Kim TWKIM@ANDREW.CMU.EDU
John Hooker JH38@ANDREW.CMU.EDU
Tepper School of Business, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Thomas Donaldson DONALDST@WHARTON.UPENN.EDU
Wharton School of Business, University of Pennsylvania
3730 Walnut Street, Philadelphia, PA 19104-6340 USA

Abstract

An important step in the development of value alignment (VA) systems in artificial intelligence (AI) is understanding how VA can reflect valid ethical principles. We propose that designers of VA systems incorporate ethics by utilizing a hybrid approach in which both ethical reasoning and empirical observation play a role. This, we argue, avoids committing "naturalistic fallacy," which is an attempt to derive "ought" from "is," and it provides a more adequate form of ethical reasoning when the fallacy is not committed. Using quantified modal logic, we precisely formulate principles derived from deontological ethics and show how they imply particular "test propositions" for any given action plan in an AI rule base. The action plan is ethical only if the test proposition is empirically true, a judgment that is made on the basis of empirical VA. This permits empirical VA to integrate seamlessly with independently justified ethical principles.

1. Introduction

Artificial intelligence (AI) technologies increasingly replace human decision makers. Worries rise about the compatibility of AI and human values. A growing number of researchers are examining how AI can acquire moral intelligence (Wallach & Allen, 2008; Burton et al., 2017; Walsh et al., 2019; Lin, Abney, & Bekey, 2011; Bringsjord, 2013; Scheutz & Arnold, 2016; Arnold, Kasenberg, & Scheutz, 2017; Arnold & Scheutz, 2018). We refer to such attempts as "value alignment" (hereafter VA). Russell, Dewey, and Tegmark (2015) highlight the need for VA and identify two options for achieving it:

> [A]ligning the values of powerful AI systems with our own values and preferences ... [could involve either] a system [that] infers the preferences of another rational or nearly rational actor by observing its behavior ... [or] could be explicitly inspired by the way humans acquire ethical values.

As this passage suggests, one option for VA is to teach machines human preferences, and another is to teach machines ethics. The word "values" in fact has this double meaning. It can refer to what humans value in the sense of what they see as subjectively preferable, or it can refer to reasonably defensible ethical principles. The distinction is important, because we acquire knowledge of the two types of values in different ways.

A similar distinction occurs in previous literature under the names *top-down* and *bottom-up* (Allen, Varner, & Zinser, 2000; Allen, 2002; Allen, Smit, & Wallach, 2005; Allen, Wallach, & Smit, 2006; Allen et al., 2005; Wallach, Allen, & Smit, 2008). Russell et al. (2015) suggest a bottom-up approach in the form of inverse reinforcement learning, which allows a machine to internalize a pattern of preferences by observing how humans actually behave (Abbeel & Ng, 2004; Ng & Russell, 2000). Reinforcement learning, and machine learning (ML) in general, offer a number of advantages but must deal with such issues as inadequate reward functions to represent complex ethical norms, biased data, and opaqueness (Arnold et al., 2017; Prince & Pinker, 1988; Marcus, 2018). A promising alternative to ML is *logic-based* VA, which has received less attention despite having a long research record (Arkoudas, Bringsjord, & Bello, 2005; Bringsjord, Arkoudas, & Bello, 2006; Bringsjord & Taylor, 2012; Bringsjord, 2017; Govindarajulu & Bringsjord, 2017; Hooker & Kim, 2018).

In this paper, we make a case for *hybrid* VA that combines ML-based and logic-based approaches. A logic-based approach is especially important because it allows the use of "independently justified" or "independently defensible" ethical principles. By these we mean principles that find their justification in ethical theory. Such principles are "normative" in the sense commonly used by moral philosophers: they are prescriptive rather than descriptive and are elements of traditional normative moral theories such as deontology, consequentialism, and virtue ethics. Such principles are increasingly discussed as candidates for computational use (Bentzen & Lindner, 2018; Lindner, Mattmüller, & Nebel, 2020; Ganascia, 2007). Independently justified principles avoid many problems, including those associated with the well-known *is-ought gap*, one aspect of which is reflected in the unconscious biases now widely studied by behavioral ethicists (Bazerman & Tenbrunsel, 2011). In turn, we propose our own version of deontological VA for use in such a hybrid approach.

After elaborating on why a purely ML-based approach is inadequate, we show how symbolic logic enables the introduction of deontological reasoning into machine ethics. Rather than opting for a particular version of moral theory, we attempt to develop a comprehensive, ecumenical framework of ethical principles (Parfit, 2011). We first articulate univeralization, utilitarian, and autonomy-based principles in the idiom of quantified modal logic. We then use these principles to derive *test propositions*, also formulated in modal logic, for each action specified by an AI rule base. The action is ethical only if the test propositions are empirically true, a judgment that can be based on machine learning and empirical VA. This permits empirical VA to integrate seamlessly with independently justified ethical principles.

2. Two Different Value Alignment Systems

Because more than one philosophical theory of mind is possible, different models of AI are are also possible, and so too, different models of VA. Broadly speaking, two categories of VA stand out, ML-based and logic-based, although neither is instantiated perfectly in any given working AI system (Table 1).

ML-based VA is connectionist. Connectionism holds that human intelligence can be explained and imitated by using artificial neural nets consisting of three kinds of connected units: input, hidden, and output (Buckner & Garson, 2019). Deep Learning (DL) exempli-

Table 1: Comparison of ML-based and logic-based VA

	ML-based	**Logic-based**
Theory of mind	Connectionism	Computationalism
Base discipline	Statistics	Logic
AI techniques	Machine learning (automated statistics, deep learning)	Symbolic AI (i.e., GOFAI: Good Old Fashioned AI)
Value alignment	Bottom-up	Top-down
Example	ML system trained by lay people's perception of fairness regarding autonomous vehicles and gender/racial discrimination	Formalized normative principles (e.g., double effect theory, categorical imperatives) using symbolic logic (e.g., quantified modal logic)
Dual process theory	System 1	System 2

fies connectionism by utilizing a complex “automated statistics” based on a large number of hidden and opaque heuristics using associations (Danks, 2014). ML’s major advantage, which is especially obvious in an end-to-end model such as DL, is its powerful ability to imitate and further strengthen skill sets in training data. DL has illustrated the power of connectionist models by capably learning human skills, especially in the domain of pattern recognition, face recognition, medical diagnostic systems, and text reading.

2.1 The Is-Ought Gap and The Problem of Bounded Ethicality

Since connectionist systems are inductive, the quality of ML-based VA relies heavily on that of inputs. If training data is biased or unethical, the system will generate well-imitated, undesirable outputs. Microsoft’s AI-based chatter-bot Tay (an acronym for “thinking about you”) was designed to engage with people on Twitter and learn from them how to carry on a conversation. When some people started tweeting racist and misogynistic expressions, Tay responded in kind. Microsoft immediately terminated the experiment (Wolf, Miller, & Grodzinsky, 2017). Most algorithmic bias problems we see now are the results of ML-based VA, which uses data sets from humans who already have implicit or explicit biases.

These mistakes reflect an error well-known to moral philosophers, the problem of deriving an “ought” from an “is,” sometimes called the “naturalistic fallacy.” From the fact that people behave in racist ways, it cannot follow that people ought to behave in such ways. While not a formal fallacy, the violation of the is-ought gap signals a form of epistemic naïveté, one that ignores the axiom in normative ethics that “no justifiable ‘ought’ can be derived directly from an ‘is”’. Disagreements about the robustness of the fallacy abound (Donaldson, 1994, 2012; Pigden, 2016; Woods & Maguire, 2017), and so this paper adopts a modest, workable interpretation of the is-ought gap coined recently by Daniel Singer, namely, “There are no valid arguments from non-normative premises to a relevantly normative conclusion” (Singer, 2015). Descriptive (or naturalistic) statements are reportive of what states of affairs are like, whereas normative statements are stipulative and action-guiding. Examples of the former are "The grass is green" and "Many people find deception to be unethical." Examples of the latter are "You ought not murder" and "Lying is unethical." Normative statements usually figure in the semantics of deontic (obligation-based) or evaluative expressions such as "ought," "impermissible," "wrong," "good," "bad," or "unethical." One may object that a high-level domain-general premise such as "machines ought to have our values no matter what" might successfully link a descriptive premise to a normative conclusion. This objection is correct, but allows the original problem to pop up again at a deeper level. "What facts," one might ask, "justify the conclusion that machines ought to imitate perfectly our behaviors?"

Using data from "unbiased" people's behaviors seems an obvious solution, but the problem is more complicated than one thinks. The preceding decade of research in behavioral ethics has shown the existence of various pernicious influences on ethical decisions, often at an unconscious level. When these influences lead to unethical behavior that conflict with an actor's moral beliefs and commitments (Moore et al., 2006), the phenomenon is often referred to as "bounded ethicality"(Bazerman, 2011; Bazerman & Tenbrunsel, 2011; Chugh, Bazerman, & Banaji, 2005; Tenbrunsel, 2005). One example of bounded ethicality is "ordinary prejudice," which reveals itself in implicit associations about gender, race, and other demographic groups (Bertrand, Chugh, & Mullainathan, 2005; Green et al., 2007; Greenwald et al., 2009; Rudman & Ashmore, 2007). These associations can lead to unintentionally discriminatory results, such as discriminatory hiring practices and unwarranted discrepancies in the evaluation of the skills and competencies of workers. Other elements of bounded ethicality include "in-group favoritism," "self-serving bias," and "motivated blindness," the last of which refers to a systemic but unconscious failure to notice unethical behavior in oneself or others even when it is in one's financial interest to do so (Bazerman & Moore, 2011; Moore, Tanlu, & Bazerman, 2010). One might consider using professional moral philosophers' opinions as training data for ML-based VA (Anderson & Anderson, 2011), but recent research shows both expert judgment generally and ethical expert judgment in particular to be frequently biased. Professional ethicists' moral intuitions and specific judgements turn out to be as vulnerable to biases or irrelevant factors as those of lay persons (Schwitzgebel & Cushman, 2012; Wiegmann, Horvath, & Meyer, 2020; Tobia, Buckwalter, & Stich, 2013; Schwitzgebel & Cushman, 2015; Egler & Ross, 2020). Because any attempt to use the ML-based VA system to generate the principles would be viciously circular, ML-based systems stand in need of independently defensible principles in order to evaluate even the training data to be used.

Logic-based VA is distinct from the ML-based in several ways. It is analogous to computationalism, in which human intelligence operates as a computer does, or in other words, in step with a set of systematic, abstract, symbol-and-rule mechanisms that are transparently expressed with formal-symbolic logic (Rescorla, 2020; Scheutz, 2002). Due to the popularity of ML systems, logic-based systems are sometimes referred to as GOFAI ("good old-fashioned AI") (Haugeland, 1985). But logic-based AI is still widely used, for instance, in the driving mechanisms of autonomous drones or cars, even though the pattern recognition mechanisms in these applications are primarily based on ML systems. Logic-based approaches are especially useful when formalizing independently defensible ethical principles of the sort invoked by professional philosophers. Such logic-based systems are sometimes labeled *symbolic AI*. Interestingly enough, formal logic is one of a few languages shared by both computer scientists and moral philosophers. Unlike eliminative (pure) connectionist systems, logic-based VA relies not on associations, but on deductive logic and logical proofs.

2.2 The Problem of System 2 and Systematicity

From a psychological perspective, ML systems are relevantly similar to what dual process theory (Kahneman, 2011) knows as "System 1" (Chauvet, 2018; Geffner, 2018; Rossi & Loreggia, 2019). It is opaque, fast, and intuitive to use. Dual process theory frames the human mind in terms of two distinctive processes: System 1 and System 2. In contrast to System 1 thinking, System 2 thinking is slow, transparent, analytical, logical, reasons-responsive, and computational. Research shows that unethical and biased decisions are correlated with System 1 thinking, and that shifting the mode to System 2 thinking is often an effective way to avoid unethical behaviors (Bazerman & Gino, 2012; Bazerman & Sezer, 2016; Zhang et al., 2015; Sezer, Gino, & Bazerman, 2015). This is despite the fact that System 1 thinking can be useful in other domains where intuitive associations are useful, such as in making heuristic decisions.

Because ML systems draw upon System 1 behavior, ML-based VA can be inherently vulnerable to unethical decision-making. The "systematicity" challenge, neglected by connectionists for decades (Calvo & Symons, 2014; Lake & Baroni, 2018; Alhama & Zuidema, 2019; Geffner, 2018; Marcus, 2001), sheds further light on this. In 1988, linguistic philosophers (Fodor & Pylyshyn, 1988) argued that connectionism confuses the intrinsically systematic nature of thought with a system of associations. More specifically, they argued that thoughts—e.g., "Mary loves John"—must involve operations with a set of rules (e.g., syntactic and semantic combinatorial relations or grammars). Pure or eliminative connectionist systems, which rely exclusively on associations, lack the ability to employ rules, and this seriously limits their ability to explain human thinking. A human who can think "Mary loves John" can also think "John loves Mary," but purely connectionist systems trained by connectionist methods cannot systematically do the latter without further resources. Responding to this challenge, many connectionists have attempted to show that structured ML systems might be redesigned, but the attempts underscore the eventual need for ML systems that employ rule-like structures.

Our purpose here is not to adjudicate this debate. However, the debate itself reveals the need for connectionist systems to be used within their legitimate scope. In that sense, our view is roughly consistent with that of Paul Smolensky who responded to the systematicity challenge in his article, "On the proper treatment of connectionism (PTC)" (1988). Similar to dual process theory, Smolensky's "proper treatment of connectionism" construes human intelligence in terms of two distinct realms: on the one hand, there is "cultural knowledge" (e.g., formalized knowledge presented by symbols and rule-like logic), and, on the other, there is "individual knowledge" (e.g., perception, intuitive processing). Connectionist systems are adequate for the latter, but not the former. The proper treatment of connectionism entails that computational systems are necessary but insufficient for language-like processing because human language operates against a backdrop of empirical, common-sensical knowledge which, in turn, allows rules themselves to make sense.

This broad point is especially relevant for moral thought, in which the "reasoning" portion of moral thinking relies upon systemic operations instead of associations. A person who can reason, "It is wrong for Jane to gratuitously lie to Mary" can also reason "It is wrong for Mary to gratuitously lie to Jane" or "It is not wrong for . . . ." Moral reasoning is fundamentally rule-based. It can be said that a person who concludes "It is wrong for Jane to lie to Mary" uses a rule such as "It is wrong for agent x to gratuitously lie to someone" and an empirical premise, "Jane gratuitously lies to Mary."[1]

3. Related Work

Bringsjord and his collaborators (Bringsjord et al., 2006; Bringsjord & Taylor, 2012; Bringsjord, 2017; Arkoudas et al., 2005; Govindarajulu & Bringsjord, 2017) are the first we know to use deontic logic to explicitly represent philosophically justifiable ethical principles such as the doctrine of double effect. Our approach dovetails with that of Bringsjord in identifying the importance of deontic logic for teaching right and wrong to machines. Since his pioneering work, many others have attempted to represent ethical principles using deontic logic. These contributions reveal the versatility of deontic logic when formalizing not only deontological moral theory but other traditions, such as areteic theory (including virtue ethics) and commandment theory.

Our work is consistent with the established deontic tradition in moral philosophy that uses deontic logic to formalize deontological moral theory. Rather than opting for a particular version of moral theory, we attempt to develop a comprehensive, ecumenical framework of ethical principles (Parfit, 2011). We offer a deontological representation of three central ethical traditions, using a generalization principle, an autonomy principle, and a deontic utility principle. Using deontology, we indicate in outline how ethical obligations can be derived from first principles instead of relying on conflicting moral intuitions of what seems fair or unbiased. While ethical philosophy has been viewed as vague and subjective by the popular imagination, the deontological approach to moral philosophy is known for offering a rigorous foundation.

Wallach et al. (2008) first suggested a hybrid approach to VA and recommended combining top-down and bottom-up approaches. Although their distinction can be more broadly construed, a typical top-down approach installs ethical principles directly into the machine, while a bottom-up approach typically asks the machine to learn prescriptive norms from experience. From an epistemological perspective, the typical bottom-up VA approach can result in teaching strategies that sometimes conflate "is" and "ought." For example, one might suggest that a machine might learn ethics through a simulated process of evolution (Conitzer et al., 2017). The fact that certain ethical norms evolve does not imply that they are valid ethical principles (Berker, 2009; Nagel, 1979; McDowell, 1995; Rachels, 1990). It is true that bottom-up approach does not automatically commit the naturalistic fallacy, particularly if ethical principles validate the norms learned in this fashion (Wallach & Allen, 2008). Nonetheless, in our approach to hybrid VA, bottom-up learning does none of the

[1] Moral particularism criticizes rule-based ethical theory, but grants easily that rules are used in moral reasoning, even as it critiques a one-size-fits-all approach. Interestingly, a rule-based or logic-based ethical theory is not committed to a rigorous one-size-fits-all approach (Smith & Dubbink, 2011).

normative work, but is used only to evaluate the truth of test propositions derived from ethical principles.

Another version of a hybrid approach to VA is advocated by Arnold et al., who argue, "architectures must explicitly represent legal, ethical and moral principles, while using them as principles for decision-making in order to achieve predictable decisions on the part of the system" and "systems that uphold those principles as much as possible represent a more ethical path than systems that are less transparent less accountably trained, and less easily corrected"(2017, p. 81). We largely agree with these authors, and our efforts are indebted to their insightful criticism of the IRL-based VA. Arnold et al. suggest that the problems in the IRL approach can be significantly addressed by an hybrid approach in which explicitly written ethical rules can be imposed as constraints on what a machine learns from observation through IRL. We follow this very path by developing deontological principles as constraints, realizing nonetheless that one must ask precisely what remains within the unconstrained space of observational learning. If what remains is learning that includes ethical norms, then once again we confront the is-ought gap. If, on the other hand, it is learning that includes empirical facts about the world, then those facts alone cannot be transformed into "oughts."

It is with this in mind that we offer a hybrid approach to VA that integrates independently justified ethical principles from the deontological tradition in ethics (Korsgaard, 1996; Nagel, 1986; O'Neill, 2014) with factual knowledge acquired through ML technology. Relevant facts may include observed preferences and values, but even such value-relevant facts cannot be the source of ethical principles.

Applying the imperative, "Thou shalt not kill," to a given action requires at a minimum that someone knows the facts relevant to the action (Hare, 1991). The relevant facts, which may include observations of human values and preferences, do not by themselves decide what is ethical, but they factor into ethical assessment. In addition, action decisions almost always take the form, "If the facts are such-and-such, then do A," which we refer to as an *action plan*. This provides a clue as to how VA can knit together empirical observation and ethical principles. The factual information in an action plan can be merged with ethical imperatives that depend on factual circumstances to arrive at an ethical judgment. The next section describes in detail how this can be accomplished.

4. Integrating Ethical Principles and Empirical VA

We now show how deontologically derived ethical principles can combine with empirical facts in a systematic way. An adequate exposition of deontological reasoning is far beyond the scope of this paper, and we do not attempt to defend the specific ethical principles we have chosen, although we briefly explain why we think they are reasonable. Relevant literature is cited for readers who wish to study the underlying arguments in detail. Our purpose here is only to show how a careful statement of ethical principles clarifies how these principles can interrelate with empirical observation in VA.

We argue that expressing ethical assertions in the idiom of quantified modal logic, as developed in Hooker and Kim (2018), makes the relationship between ethical principles and empirical observation perspicuous. Specifically: *ethical principles imply certain logical propositions that must be true in order for a given action plan to be ethical, and empir-*

ical observation determines whether these propositions are, in fact, true. We refer these propositions as *test propositions*, whose empirical evaluation typically requires observation of human values, beliefs, and behavior. The test propositions need not appear alongside the action plans in an AI system, but they can be generated and evaluated automatically if desired (Section 4.5).

Thus the role of ethics in hybrid VA is to derive necessary conditions for the rightness of specific actions, and the role of empirical VA is to ascertain whether these conditions are satisfied in the real world.

4.1 Actions and Reasons

Deontology derives ethical principles from the logical structure of action (Kant, 1785; Wood, 1999; O'Neill, 2014; Hooker & Kim, 2019). It begins with the necessity of distinguishing free action from mere behavior, insofar as causally speaking, both are determined by chemical and physical forces. Contemporary deontological thinkers usually base the distinction between free and causally determined behavior on a Kantian *dual standpoint* theory of ethics that identifies free action as behavior for which the agent has *reasons* (Bilgrami, 1996; Korsgaard, 1996; Nagel, 1986; Nelkin, 2000). Such reasons are not themselves psychological causes or motivations, but considerations that the agent consciously makes to justify a choice. The reasons need not be good or convincing ones from another agent's perspective, but must be sufficiently coherent to serve as an explanation of why the agent chose the action.

Ethical principles are necessary conditions for the coherence or intelligibility of the reasons behind an action. While a number of necessary conditions for coherence are possible, ethical principles rest on the *universality of reason*: an agent who takes a set of reasons as justifying an action must, in order to be consistent take the reasons as justifying the same action for any agent to whom those reasons apply.

We focus on the three ethical principles that have been most intensely studied in the literature—generalization, utility maximization, and respect for autonomy. Each states a necessary condition for ethical conduct. We make no claim that they are exhaustive, but only that they illustrate how empirical VA can be anchored by ethical principles.

Before proceeding, two caveats are in order. First, in this paper we do not attempt to convince readers of the superiority of the deontological tradition or its premise that principles can be discovered through an analysis of the logical structure of action. Our aim is more modest: to show that deontology is particularly suitable for hybrid VA. Two of the three principles we employ, generalization and respect for autonomy, have historical roots in Kant's The Formula of the Universal Law and The Formula of Humanity (Wood, 1999), although our formulations of them differ. Second, we also use a deontic model of utilitarianism (Cummiskey, 1996) in order to make utilitarianism consistent with the other two other principles.

4.2 Generalization Principle

The universality of reason leads immediately to the *generalization principle*: a rational agent must believe that his/her reasons for acting are consistent with the assumption that all rational agents to whom the reasons apply could engage in the same actions (O'Neill, 2014; Wood, 1999).

As an example, suppose I see wristwatches on open display in a shop and steal one. My reasons for the theft are that I would like to have a new watch, and that I can get away with taking one.[2] At the same time, I cannot rationally believe that I would be able to get away with the theft if *everyone* stole watches when these reasons apply. The shop would install security measures to prevent theft, which is inconsistent with one of my reasons for stealing the watch. The theft therefore violates the generalization principle.

To give these ideas more precision, we express the action plan and generalization principle in the language of quantified modal logic. In so doing, we do not define a deductive system or propose formal semantics, as they are unnecessary for our project. We merely borrow logical notation in order to allow a more rigorous formulation and application of ethical principles.

The decision to steal a watch can be expressed in logical notation as follows. Define predicates:

$$\begin{aligned} C_1(a) &= \text{Agent } a \text{ would like to possess an item on} \\ &\ \text{display in a shop.} \\ C_2(a) &= \text{Agent } a \text{ can get away with stealing the item.} \\ A_1(a) &= \text{Agent } a \text{ will steal the item.} \end{aligned}$$

Because the agent's reasons are an essential part of moral assessment, we evaluate the agent's *action plan*, which states that the agent will take a certain action when certain reasons apply. In this case, the action plan is:

$$\big(C_1(a) \wedge C_2(a)\big) \Rightarrow_a A_1(a) \tag{1}$$

Here, $\Rightarrow_a$ is not logical entailment but indicates that agent a regards $C_1(a)$ and $C_2(a)$ as justifying $A_1(a)$. The reasons in the action plan should be the most general set of conditions that the agent takes as justifying the action. Thus the action plan refers to an item in a shop rather than specifically to a watch, because the fact that it is a watch is not relevant to the justification; what matters is whether the agent wants the item and can get away with stealing it.

We can now state the generalization principle using quantified modal logic. Let $C(a) \Rightarrow_a A(a)$ be an action plan for agent a, where $C(a)$ is a conjunction of the reasons for taking action $A(a)$. The action plan is generalizable if and only if:

$$\Diamond_a P\Big(\forall x\big(C(x) \Rightarrow_x A(x)\big) \wedge C(a) \wedge A(a)\Big) \tag{2}$$

Here, $P(S)$ means that it is physically possible for proposition S to be true, and $\Diamond_a S$ means that a can rationally believe S. The proposition $\Diamond_a S$ is equivalent to $\neg\Box_a\neg S$, where $\Box_a\neg S$ means that rationality requires require a to deny S.[3] Thus (2) says that agent a can

[2] In practice, the reasons for theft are likely to be more complicated than this. I may be willing to steal partly because I believe the shop can easily withstand the loss, no employee will be disciplined or terminated due to the loss, I will not feel guilty afterward, and so forth. But for purposes of illustration we suppose there are only two reasons.

[3] The operators $\Diamond$ and $\Box$ have a somewhat different interpretation here than in traditional epistemic and doxastic modal logics, but the identity $\Diamond S \equiv \neg\Box\neg S$ holds as usual.

rationally believe that it is possible for everyone to have the same action plan as a, even while a's reasons still apply and a takes the action.

Returning to the theft example, the condition (2) becomes the test proposition for action plan (2):

$$\Diamond_a P\Big(\forall x\big(C_1(x) \wedge C_2(x) \Rightarrow_x A_1(x)\big) \wedge C_1(a) \wedge C_2(a) \wedge A_1(a)\Big) \tag{3}$$

This says that it is rational for a to believe that it is physically possible for the following to be true simultaneously: (a) everyone steals when the stated conditions apply, and (b) the conditions apply and a steals. Since (3) is false, action plan (1) is unethical.

The necessity of (3) for the rightness of action plan (1) is anchored in deontological theory, while the falsehood of (3) is a fact about the world. This fact might be inferred by collecting responses from shop owners about how they would react if theft were widespread. *Thus ethics and empirical VA work together in a very specific way: ethics tells us that the test proposition (3) must be true if the theft is to be ethical, and empirical VA provides evidence that bears on whether (3) is true.*

An action plan in the autonomous vehicle domain might be:

$$C_3(a) \Rightarrow_a A_2(a) \tag{4}$$

where

$C_3(a)$ = An ambulance under the control of agent a can reach its destination sooner by using siren and lights.
$A_2(a)$ = Agent a will direct an ambulance to use siren and lights.

Agent a is the ambulance driver, or in the case of an autonomous vehicle, the designer of the software that controls the ambulance. The generalization principle yields the test proposition:

$$\Diamond_a P\Big(\forall x\big(C_3(x) \Rightarrow_y A_2(x)\big) \wedge C_3(a) \wedge A_2(a)\Big) \tag{5}$$

This says that it is rational for agent a to believe that siren and lights could continue to hasten arrival if all ambulances used them for all trips, emergencies and otherwise. If empirical VA reveals that most drivers would ignore siren and lights if they were universally abused in this fashion, then we have evidence that (5) is false, in which case action plan (4) is unethical.

4.3 Maximizing Utility

Utilitarianism is normally understood as a *consequentialist* theory that evaluates an act by its actual consequences. Specifically, an act is ethical only if it maximizes total net expected utility across all who are affected. Yet the utilitarian principle can also be construed in a deontological fashion (Cummiskey, 1996), which allows it to be interpreted as requiring the agent to select actions that the agent can rationally believe will maximize utility. While utilitarians frequently view utility maximization as the sole ethical principle, it can be seen as an additional necessary condition for an ethical action. The other non-utilitarian principles remain in force because only actions that satisfy the other principles are considered options for maximizing utility.

In a deontological analysis, utility is not what people generally value but what the agent is rationally committed to valuing. The logic of means and ends requires that the agent regard some end as *intrinsically* valuable (such as happiness), and the universality of reason requires that it be seen as valuable for any agent. A utilitarian believes this commits the agent to selecting actions that maximize the expected net sum of utility over everyone who is affected.[4]

The utilitarian principle can be formalized by requiring that a given action plan create at least as much utility as any other available action plan. Let $u(C(a), A(a))$ be a utility function that measures the total net expected utility of action $A(a)$ under conditions $C(a)$. Then an action plan $C(a) \Rightarrow_a A(a)$ satisfies the utilitarian principle only if agent a can rationally believe that action $A(a)$ creates at least as much utility as any ethical action that is available under the same circumstances. This can be written:

$$\Diamond_a \forall A' \Big(E\big(C(a), A'(a)\big) \rightarrow u\big(C(a), A(a)\big) \geq u\big(C(a), A'(a)\big) \Big) \tag{6}$$

where A' ranges over actions. The predicate $E(C(a), A'(a))$ means that action $A'(a)$ is available for agent a under conditions $C(a)$, and that the action plan $C(a) \Rightarrow_a A'(a)$ is generalizable and respects autonomy.[5] Note that we are now quantifying over predicates and have therefore moved into second-order logic.

Popular views about acceptable behavior frequently play a role in applications of the utilitarian principle. For example, in some parts of the world, drivers consider it wrong to enter a stream of moving traffic from a side street without waiting for a gap in the traffic. In other parts of the world this can be acceptable, because drivers in the main thoroughfare expect it and make allowances. Suppose driver a's action plan is $(C_4(a) \wedge C_5(a)) \Rightarrow_a A_3(a)$, where:

$C_4(a) =$ Driver a wishes to enter a main thoroughfare.
$C_5(a) =$ Driver a can enter a main thoroughfare by moving into the traffic without waiting for a gap.
$A_3(a) =$ Driver a will move into traffic without waiting for a gap.

As before, driver a is the designer of the software if the vehicle is autonomous. Using (6), the driver's action plan maximizes utility only if the following test proposition is true:

$$\begin{aligned} &\Diamond_a \forall A' \Big(E\big(C_4(a), C_5(a), A'(a)\big) \rightarrow \\ &\qquad\qquad u\big(C_4(a), C_5(a), A_3(a)\big) \geq u\big(C_4(a), C_5(a), A'(a)\big) \Big) \end{aligned} \tag{7}$$

Suppose we wish to design driving policy in a context where pulling immediately into traffic is considered unacceptable. Then, doing so is a dangerous move that no one is expecting and an accident could result. Waiting for a gap in the traffic results in greater net expected utility, or formally, $u(C_4(a), C_5(a), A_3(a)) < u(C_4(a), C_5(a), A_4(a))$, where $A_4(a)$ is the action of moving into traffic after waiting for a gap. So (7) is false, and its falsehood

[4] Alternatively, one might argue that maximizing the minimum utility over those affected (or achieving a lexicographic maximum) is the rational way to take everyone's utility into account, after the fashion of John Rawls's difference principle (Rawls, 1971). Or one might argue for some rational combination of utilitarian and equity objectives (Karsu & Morton, 2015; Hooker & Williams, 2012). However, for many practical applications, simple utility maximization appears to be a sufficiently close approximation to a "rational" choice, and to simplify exposition we assume so in this paper.

[5] For "respecting autonomy," see the next section.

can be inferred by collecting popular views about acceptable driving behavior. Observed preferences and values are therefore relevant to an ethical assessment, but they alone do not determine the assessment.

Again, we have a clear demonstration of how ethical principles can combine with empirical VA. The utilitarian principle tells us that a particular action plan is ethical only if test proposition (7) is true, and empirical VA tells us whether (7) is true.

A similar approach can accommodate other situations in which popular expectations bear on ethical decisions. For example, it has been observed that people may expect different ethical norms to be followed by machine agents rather than by humans (Malle et al., 2015). This could affect generalizability as well as a utilitarian assessment, because there may be different implied promises or agreements concerning machines than humans. Yet again, expectations alone do not determine the ethical outcome.

4.4 Respect for Autonomy

A third ethical principle requires agents to respect the autonomy of other agents. Specifically, an agent should not adopt an action plan that the agent is rationally constrained to believe is inconsistent with an ethical action plan of another agent, without informed consent. Murder, enslavement, and inflicting serious injury are extreme examples of autonomy violations because they interfere with many ethical action plans. Coercion may or may not violate autonomy, depending on precisely how action plans are formulated.[6]

The argument for respecting autonomy is basically as follows. Suppose I violate someone's autonomy for certain 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. To avoid contradicting myself, I must avoid interfering with other action plans.

To formulate an autonomy principle, we say that agent a's action plan $C(a) \Rightarrow_a A(a)$ is consistent with b's action plan $C'(b) \Rightarrow_b A'(b)$ when:

$$\Diamond_a P\big(A(a) \wedge A'(b)\big) \ \vee \ \neg\Box_a P\big(C(a) \wedge C'(b)\big) \tag{8}$$

This says that agent a can rationally believe that the two actions are mutually consistent, or can rationally believe that the reasons for the actions are mutually inconsistent. The latter suffices to avoid inconsistency of the action plans, because if the reasons for them cannot both apply, the actions can never come into conflict.

As an example of how coercion need not violate autonomy, suppose agent b wishes to catch a bus and has decided to cross the street to a bus stop (provided no traffic is coming). The agent's action plan is

$$\big(C_6(b) \wedge C_7(b) \wedge \neg C_8(b)\big) \Rightarrow_b A_5(b) \tag{9}$$

[6] A more adequate analysis leads to a principle of *joint* autonomy, according to which it is violation of autonomy to adopt an action plan that is mutually inconsistent with action plans of a set of other agents, when those other action plans are themselves mutually consistent. Joint autonomy addresses situations in which an action necessarily interferes with the action plan of some agent but no particular agent, as when someone throws a bomb into a crowd. A general formulation of the joint autonomy principle in terms of modal operators is given in Hooker and Kim (2018). This and other complications are discussed in Hooker (2018).

where

$C_6(b)$ = Agent b wishes to catch a bus.
$C_7(b)$ = There is a bus stop across the street from b.
$C_8(b)$ = There are cars approaching b.
$A_5(b)$ = Agent b will cross the street.

Agent a sees agent b begin to cross the street and forcibly pulls b out of the path of an oncoming car that b does not notice. Agent a's action plan is:

$$\big(C_8(b) \wedge C_9(b)\big) \Rightarrow_a A_6(a,b) \tag{10}$$

where

$C_9(b)$ = Agent b is about to cross the street.
$A_6(a,b)$ = Agent a will prevent agent b from crossing the street.

Agent a does not violate agent b's autonomy, even though there is coercion. Their action plans (9) and (10) are consistent with each other, because the condition (8) yields the test proposition:

$$\Diamond_a P\big(A_5(b) \wedge A_6(a,b)\big) \ \vee \ \neg\Box_a P\big(C_6(b) \wedge C_7(b) \wedge \neg C_8(b) \wedge C_8(b) \wedge C_9(b)\big) \tag{11}$$

This means that either (a) agent a can rationally believe that the two actions are consistent with each other, or (b) agent a can rationally believe that the antecedents of (9) and (10) are mutually inconsistent. As it happens, the two actions are obviously not consistent with each other, and so (a) is false. However, agent a can rationally believe that the antecedents of (9) and (10) are mutually inconsistent, because $C_8(b)$ and $\neg C_8(b)$ are contradictory. This means (b) is true, which implies that condition (11) is satisfied, and there is no violation of autonomy.

Again, this clearly distinguishes the roles of ethics and empirical observation in VA. Ethical reasoning tells us that the test proposition (11) must be true if autonomy is to be respected, whereas observation of the world tells us whether (11) is true.

In saying that coercion can be ethical, we do not imply that a violation of autonomy can be ethical. Coercion must be consistent with the coerced agent's action plan, as in the above example. Coercion can also be ethical when there is implied or informed consent, or when it is necessary to prevent unethical behavior (as in self-defense).[7] Interfering with an unethical action plan is no violation of autonomy because an unethical action plan is, strictly speaking, not an action plan due to the absence of a coherent set of reasons for undertaking it. An action plan is considered unethical in this context when it violates the generalization or utility principle, or interferes with an action plan that does not violate one of these principles, and so on recursively. Thus, coercion is ethical in an act of self-defense, or to stop someone from unethically harming others.

To illustrate how autonomy may play a role in the ethics of driving, suppose that a pedestrian b dashes in front of a's rapidly moving car. Driver a can slam on the brake and

[7] Coercion can be ethical when there is informed consent to a risk of interference, because giving informed consent is equivalent to including the possibility of interference as one of the antecedents of the action plan. This occurs, for example, when a medical test subject gives consent with the knowledge that an experimental drug may cause illness, even though administering a drug that turns out to be harmful is a form of coercion.

avoid impact with the pedestrian, but another driver c is following closely and a sudden stop could cause a crash. The driver a must choose between two possible action plans:

$$\big(C_{10}(a,b) \wedge C_{11}(a,c)\big) \Rightarrow_a A_7(a) \tag{12}$$

$$\big(C_{10}(a,b) \wedge C_{11}(a,c)\big) \Rightarrow_a \neg A_7(a) \tag{13}$$

where

$C_{10}(a,b) =$ Pedestrian b is dashing in front of a's car.
$C_{11}(a,c) =$ Driver c is closely following a's car.
$A_7(a) =$ Agent a will immediately slam on the brake.

Meanwhile, the pedestrian b has any number of action plans that are clearly inconsistent with death or serious injury. Let $C_{12}(b) \Rightarrow_b A_8(b)$ be one of them. Also, driver c of the other car (there is only one occupant) has action plans that are inconsistent with an injury. We suppose that $C_{13}(c) \Rightarrow_c A_9(c)$ is one of them.

We first check whether hitting the brakes, as in action plan (12), is inconsistent with the other driver's action plan $C_{13}(c) \Rightarrow_c A_9(c)$. The test proposition is

$$\Diamond_a P\big(A_7(a) \wedge A_9(c)\big) \;\vee\; \neg\Box_a P\big(C_{10}(a,b) \wedge C_{11}(a,c) \wedge C_{13}(c)\big) \tag{14}$$

The first disjunct is clearly true, because a can rationally believe that it is *possible* that hitting the brake is consistent with avoiding a rear-end collision and therefore with any planned action $C_{13}(c) \Rightarrow_c A_9(c)$, even if this is improbable. So action plan (12) does not violate joint autonomy.

We now check whether a failure to hit the brake, as in action plan (13), is inconsistent with the pedestrian's action plan $C_{12}(b) \Rightarrow_b A_8(b)$. There is no violation of autonomy if

$$\Diamond_a P\big(\neg A_7(a) \wedge A_8(b)\big) \;\vee\; \neg\Box_a P\Big(C_{10}(a,b) \wedge C_{11}(a) \wedge C_{12}(b)\Big) \tag{15}$$

The first disjunct of (15) is clearly false for b's action plan $C_{12}(b) \Rightarrow_b A_8(b)$, because driver a cannot rationally believe that a failure to hit the brake is consistent with it. The second disjunct is likewise false, because driver a has no reason to believe that $C_{10}(a,b)$, $C_{11}(a,c)$, and $C_{12}(b)$ are mutually inconsistent. Thus (15) is false, and we have a violation of autonomy. The driver should therefore slam on the brakes. There is no need to check the other ethical principles, because only one of the possible action plans satisfies the autonomy principle.

4.5 Implementation Issues

While it is not our purpose to address engineering aspects of deontically-grounded VA, we can take note of some implementation issues that arise. The main implication of our proposal is that the portion of an AI system that makes ethically relevant decisions must be rule-based (i.e., an instance of GOFAI) because it must consist of action plans. Fortuitously, action plans have an if–then structure that is convenient for coding rules.

One can ask whether a rule-based system is adequate for the complexities of real-life decision-making, but this is, of course, a problem that is not confined to deontically-based VA. We do not attempt here to judge the versatility of rule-based AI, but we note that it seems to be increasingly viewed as technically viable and even necessary due to the nontransparency of deep learning and support vector machines. Regarding autonomous vehicles, for example, Brandom (2018) states: "Many companies have shifted to rule-based AI, an older technique that lets engineers hard-code specific behaviors or logic into an otherwise self-directed system." The technical community has ample experience at accurately coding and debugging huge rule-based systems. An ordinary (non-self-driving) automobile is already regulated by more than 100,000 lines of code. Ethics-based systems can evolve through several versions and be updated as necessary, as with any other type of complex software. Rule-based AI can also be combined with machine learning (Woźniak & Połap, 2020). Even in a pure ML system, it is possible to derive rules that approximate the directives generated by ML (Soares, Angelov, & Costa, 2020) and perhaps subject them to ethical evaluation.

The test propositions used to evaluate the ethical status of action plans need not appear in the AI rule base, and it is a further implementation decision whether to generate them automatically. This is fairly straightforward (less so for the utilitarian test), because the procedure for doing so can be clearly specified as shown above. Machine learning and other forms of empirical VA can then be used to evaluate the truth of the test propositions.

5. Conclusion

Humanity's goal should be to invest machines with a moral sensitivity that mimics the human conscience. But conscience is dynamic rather than static, and adjusts ethical principles systematically to empirical observations. In this paper we have elaborated two challenges to AI moral reasoning that spring from the interrelation of facts and values. The first is a confusion that mistakenly identifies facts for values; the second is a confusion that misunderstands the process of moral reasoning. In addressing these challenges, we have identified how and why AI can commit the naturalistic fallacy, move illicitly from "is's" to "oughts," and oversimplify the process of moral reasoning. We have sketched, in response, a proposal for understanding moral reasoning in machines, one that highlights how deontological ethical principles can interact with factual states of affairs.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*.

Alhama, R. G., & Zuidema, W. (2019). A review of computational models of basic rule learning: The neural-symbolic debate and beyond. *Psychonomic Bulletin & Review*, *26*(4) 1174–1194.

Allen, C. (2002). Calculated morality: Ethical computing in the limit. *Cognitive, Emotive and Ethical Aspects of Decision Making and Human Action*, *1*, 19–23.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*(3), 149–155.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics?. *IEEE Intelligent Systems*, *21*(4), 12–17.

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 Anderson, M., & Anderson, S. L. (Eds.), *Machine Ethics* pp. 476–492 Cambridge University Press.

Arkoudas, K., Bringsjord, S., & Bello, P. (2005). Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*, 17–23.

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment–What will keep systems accountable?. In *3rd International Workshop on AI, Ethics, and Society*.

Arnold, T., & Scheutz, M. (2018). The "big red button" is too late: An alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology*, *20*(1), 59–69.

Bazerman, M. H. (2011). Bounded ethicality in negotiations. *Negotiation and Conflict Management Research*, *4*(1), 8–11.

Bazerman, M. H., & Gino, F. (2012). Behavioral ethics: Toward a deeper understanding of moral judgment and dishonesty. *Annual Review of Law and Social Science*, *8*, 85–104.

Bazerman, M. H., & Moore, D. (2011). Is it time for auditor independence yet?. *Accounting, Organizations and Society*, *36*(4-5), 310–312.

Bazerman, M. H., & Sezer, O. (2016). Bounded awareness: Implications for ethical decision making. *Organizational Behavior and Human Decision Processes*, *136*, 95–105.

Bazerman, M. H., & Tenbrunsel, A. E. (2011). *Blind spots: Why we fail to do what's right and what to do about it*. Princeton University Press.

Bentzen, M. M., & Lindner, F. (2018). A formalization of Kant's second formulation of the categorical imperative.. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM)*.

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs*, *37*(4), 293–329.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit discrimination. *American Economic Review*, *95*(2), 94–98.

Bilgrami, A. (1996). *Self-knowledge and resentment*. Harvard University Press.

Brandom, R. (2018). Self-driving cars are headed toward an AI roadblock. *The Verge*.

Bringsjord, S. (2013). *What robots can and can't be*, Vol. 12. Springer Science & Business Media.

Bringsjord, S. (2017). A 21st-century ethical hierarchy for robots and persons. In *A world with robots*, pp. 47–61 Springer.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems 21*(4) 38–44.

Bringsjord, S., & Taylor, J. (2012). The divine-command approach to robot ethics. *Robot Ethics: The Ethical and Social Implications of Robotics*, 85–108 MIT Press.

Buckner, C., & Garson, J. (2019). Connectionism. In Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy* Metaphysics Research Lab, Stanford University.

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical considerations in artificial intelligence courses. *AI magazine*, *38*(2) 22–34.

Calvo, P., & Symons, J. (2014). *The architecture of cognition: Rethinking Fodor and Pylyshyn's systematicity challenge*. MIT Press.

Chauvet, J.-M. (2018). The 30-year cycle in the AI debate. *arXiv preprint arXiv:1810.04053*.

Chugh, D., Bazerman, M. H., & Banaji, M. R. (2005). Bounded ethicality as a psychological barrier to recognizing conflicts of interest. In Moore, D. A., Cain, D. M., Loewenstein, G., & Bazerman, M. H. (Eds.), *Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy* pp. 74–95. Cambridge University Press.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *Proceedings of 31st AAAI Conference on Artificial Intelligence*, pp. 4831–4835.

Cummiskey, D. (1996). *Kantian consequentialism*. Oxford University Press.

Danks, D. (2014). Learning. In Keith, F., & Ramsey, W. M. (Eds.), *Cambridge handbook to artificial intelligence*, 151–167 Cambridge University Press.

Donaldson, T. (1994). When integration fails: The logic of prescription and description in business ethics. *Business Ethics Quarterly*, *4*(2), 157–169.

Donaldson, T. (2012). The epistemic fault line in corporate governance. *Academy of Management Review*, *37*(2), 256–271.

Egler, M., & Ross, L. D. (2020). Philosophical expertise under the microscope. *Synthese*, *197*(3), 1077–1098.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Ganascia, J.-G. (2007). Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, *9*(1), 39–47.

Geffner, H. (2018). Model-free, model-based, and general intelligence. *arXiv preprint arXiv:1806.02308*.

Govindarajulu, N. S., & Bringsjord, S. (2017). On automating the doctrine of double effect. *arXiv preprint arXiv:1703.08922*.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of General Internal Medicine*, *22*(9), 1231–1238.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity.. *Journal of Personality and Social Psychology*, *97*(1), 17.

Hare, R. M. (1991). *The language of morals.* Oxford Paperbacks.

Haugeland, J. (1985). *Artificial Intelligence: The very idea.* MIT Press.

Hooker, J. N., & Kim, T. W. (2019). Truly autonomous machines are ethical. *AI Magazine*, *40*(4).

Hooker, J. N., & Kim, T. W. N. (2018). Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '18) 130–136.

Hooker, J. N., & Williams, H. P. (2012). Combining equity and utilitarianism in a mathematical programming model. *Management Science*, *58*(9), 1682–1693.

Hooker, J. (2018). *Taking ethics seriously: Why ethics is an essential tool for the modern workplace.* Taylor & Francis.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

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

Karsu, Ö., & Morton, A. (2015). Inequality averse optimization in operational research. *European Journal of Operational Research*, *245*, 343–359.

Korsgaard, C. M. (1996). *The sources of normativity.* Cambridge University Press.

Lake, B., & Baroni, M. (2018). Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *ArXiv preprint 1711.00350.*

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence*, *175*(5-6), 942.

Lindner, F., Mattmüller, R., & Nebel, B. (2020). Evaluation of the moral permissibility of action plans. *Artificial Intelligence*, 287.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124.

Marcus, G. (2001). *The algebraic mind.* MIT Press.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint 1801.00631.*

McDowell, J. (1995). Two sorts of naturalism. In R., H., G., L., & W., Q. (Eds.), *Virtues and reasons: Philippa Foot and moral theory; Essays in Honour of Philippa Foot.* Clarendon Press.

Moore, D. A., Tanlu, L., & Bazerman, M. H. (2010). Conflict of interest and the intrusion of bias. *Judgment and Decision Making*, *5*(1), 37.

Moore, D. A., Tetlock, P. E., Tanlu, L., & Bazerman, M. H. (2006). Conflicts of interest and the case of auditor independence: Moral seduction and strategic issue cycling. *Academy of Management Review*, *31*(1), 10–29.

Nagel, T. (1979). Ethics without biology. In Nagel, T. (Ed.), *Mortal questions* 142–46. Cambridge University Press.

Nagel, T. (1986). *The view from nowhere.* Oxford University Press, Oxford.

Nelkin, D. K. (2000). Two standpoints and the belief in freedom. *Journal of Philosophy*, *97*, 564–576.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings 17th International Conference on Machine Learning*, 663–670. Morgan Kaufmann.

O'Neill, O. (2014). *Acting on principle: An essay on Kantian ethics (2nd ed.)* Cambridge University Press, Cambridge, UK.

Parfit, D. (2011). *On what matters*, Vol. 1. Oxford University Press.

Pigden, C. (2016). Hume on *is* and *ought*: Logic, promises, and the Duke of Wellington. In *The Oxford handbook of Hume* 401–415. Oxford University Press.

Prince, A., & Pinker, S. (1988). Rules and connections in human language. *Trends in Neurosciences*, *11*(5), 195–202.

Rachels, J. (1990). *Created from animals.* Oxford University Press.

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

Rescorla, M. (2020). The computational theory of mind. In Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Rossi, F., & Loreggia, A. (2019). Preferences and ethical priorities: Thinking fast and slow in AI. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 3–4.

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the implicit association test. *Group Processes & Intergroup Relations*, *10*(3), 359–372.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, *36*(4), 105–114.

Scheutz, M. (2002). Computationalism—The next generation. In *Computationalism: New Directions*, 1–21, MIT Press.

Scheutz, M., & Arnold, T. (2016). Are we ready for sex robots?. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 351–358.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, *27*(2), 135–153.

Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, *141*, 127–137.

Sezer, O., Gino, F., & Bazerman, M. H. (2015). Ethical blind spots: Explaining unintentional unethical behavior. *Current Opinion in Psychology*, *6*, 77–81.

Singer, D. J. (2015). Mind the is-ought gap. *The Journal of Philosophy*, *112*(4), 193–210.

Smith, J., & Dubbink, W. (2011). Understanding the role of moral principles in business ethics: A Kantian perspective. *Business Ethics Quarterly*, *21*(2), 205–231.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23.

Soares, E. A., Angelov, P. P., & Costa, B. (2020). Explaining deep learning models through rule-based approximation and visualization. *IEEE Transactions on Fuzzy Systems*.

Tenbrunsel, A. E. (2005). Commentary: Bounded ethicality and conflicts of interest. In Moore, D. A., Cain, D. M., Loewenstein, G., & Bazerman, M. H. (Eds.), *Conflicts of Interest: Challenges and Solutions in Business, Law, Medicine, and Public Policy*, p. 96. Cambridge University Press.

Tobia, K., Buckwalter, W., & Stich, S. (2013). Moral intuitions: Are philosophers experts?. *Philosophical Psychology*, *26*(5), 629–638.

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

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong.* Oxford University Press.

Walsh, T., Levy, N., Bell, G., Elliott, A., Maclaurin, J., Mareels, I., & Wood, F. (2019). *The effective and ethical development of artificial intelligence: An opportunity to improve our wellbeing.* Australian Council of Learned Academies.

Wiegmann, A., Horvath, J., & Meyer, K. (2020). Intuitive expertise and irrelevant options. *Oxford Studies in Experimental Philosophy*, *3* (3), 275.

Wolf, M. J., Miller, K. W., & Grodzinsky, F. S. (2017). Why we should have seen that coming: Comments on Microsoft's Tay "experiment," and wider implications. *The ORBIT Journal*, *1*(2), 1–12.

Wood, A. W. (1999). *Kant's ethical thought.* Cambridge University Press.

Woods, J., & Maguire, B. (2017). Model theory, Hume's dictum, and the priority of ethical theory. *Ergo*, *14*(4).

Woźniak, M., & Połap, D. (2020). Intelligent home systems for ubiquitous user support by using neural networks and rule-based approach. *IEEE Transactions on Industrial Informatics*, *16*(4), 2651–2658.

Zhang, T., Fletcher, P. O., Gino, F., & Bazerman, M. H. (2015). Reducing bounded ethicality: How to help individuals notice and avoid unethical behavior. *Organizational Dynamics*, *44*(4), 310–317.