Value Alignment or Misalignment – What Will Keep Systems Accountable?

Thomas Arnold and **Daniel Kasenberg** and **Matthias Scheutz**
Department of Computer Science
Tufts University
Medford, MA 02155

Abstract

Machine learning's advances have led to new ideas about the feasibility and importance of machine ethics keeping pace, with increasing emphasis on safety, containment, and alignment. This paper addresses a recent suggestion that *inverse reinforcement learning* (IRL) could be a means to so-called "value alignment." We critically consider how such an approach can engage the social, norm-infused nature of ethical action and outline several features of ethical appraisal that go beyond simple models of behavior, including unavoidably temporal dimensions of norms and counterfactuals. We propose that a hybrid approach for computational architectures still offers the most promising avenue for machines acting in an ethical fashion.

Introduction

Machine learning has attracted attention for widening the scope of plausibly successful applications (not least, in the public eye, from beating the world champion of Go for the first time (Silver et al. 2016; Borowiec 2016)). Having brought some kinds of tasks, especially game-playing, from the distant computational horizon into present-day achievement, it was only natural that fields or abilities formerly thought exclusively "human" would be tackled by a machine learning approach. Even as examples of artificial intelligence have brought serious public scrutiny to their societal impacts (e.g., the *White House Initiative on AI*) and troubling features "in the wild," many find promise in machine learning to ensure autonomous systems reliably do what they should or should not do. Russell et al. (Russell, Dewey, and Tegmark 2016) have notably suggested "value alignment" as a worthwhile target to reach via inverse reinforcement learning (IRL). The IRL approach would work to train artificial intelligence to behave as human beings wish, including ethically appropriate behavior. AI "must do what we want it to do," and inverse reinforcement learning, Russell implies, could provide the needed training to do so, albeit with some minor "gaps" here and there (Wolchover 2015).

In this paper we argue that the proposal to achieve value alignment through inverse reinforcement learning is ethically inadequate for a computational architecture designed to guide artificial agents. We explain how the social character of ethical evaluation and practical reasoning puts a substantive burden on such agents to account for what they are doing. We outline a number of technical challenges to reinforcement learning, including data bias, generalization issues, and the adequacy of reward functions to represent temporally-complex norms. We conclude that 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. Systems that uphold those principles as much as possible represent a more ethical path than systems less transparent, less accountably trained, and less easily corrected.

Background and Motivation

The relationship between artificial intelligence and ethics has often taken the shape of "machine ethics," a project of "adding" some form of ethics to a system's decision-making procedures. (Anderson and Anderson 2011). Approaches to machine ethics have varied in terms of their reliance on logic, training through modeled behavior, and the actual ethical content being promoted – whether leaning on deontic logic (Bringsjord, Arkoudas, and Bello 2006), analogical reasoning (Dehghani et al. 2008; Blass and Forbus 2015), or neural networks representing motivations (Sun 2013) – but several architectures have sought to account for how that ethical "addition" would find its way into the system. With robots especially, that effort has entailed asking what ethical theory (deontological, utilitarian, virtue, particular religious traditions), or even metaethics, defines what values a system would have (Abney 2012; Bekey et al. 2012). On the performance side, there have been ongoing questions, but few spelled-out programs, for what evaluation or testing of an autonomous system's ethics could look like (e.g., the idea of a "moral Turing Test" (Allen, Smit, and Wallach 2005; Arnold and Scheutz 2016)).

With the burgeoning role of machine learning across different domains, these questions have taken on a different arrangement of priorities. The idea of training a system on data (either supervised or unsupervised) has captured more and more attention as a way to understand how ethics and AI might best function in concert. Coding ethical values "by hand," in the manner of many other traditional forms of coding – seems destined for the lesser task of lending

basic "scaffolding" within which machine learning can operate (Tanz 2016). While this trend has itself elicited some alarmed critiques of AI bias and systemic blind spots, there have been efforts to train for ethical values through richer kinds of material (e.g., narrative (Riedl and Harrison 2016)).

In general, the movement toward machine learning has contributed to a turn away from "machine ethics" typically construed (especially any top-down articulations of rules, norms, or model behavior) and toward ensuring that an AI system's learning and growth does not turn against human interests. Imbuing a system with ethics throughout, in other words, has given way to constraining the system from the outside: making sure artifical intelligence does not escape its confines to damaging effect. Without the boundaries provided by a game like chess or Go, AI systems are projected to be possible existential threats: with increased power, they could learn to manipulate their own learning and control functions (Bostrom 2014; Bostrom and Yudkowsky 2014). There has been pushback against overhyping AI advancements, such as the idea of an impending singularity (Walsh 2016), but the risk assessment of AI systems as threats to develop beyond control (Yudkowsky 2008) has underwritten the recent launching of research programs of "alignment," "safety engineering," or containment (Taylor et al. 2016). The project of machine ethics, from the vantage of many of these projects, is subordinate to the longer-term need for confined, safe, self-improving systems (Yampolskiy 2013).

Value Alignment and Inverse Reinforcement Learning

Russell et al. (Russell, Dewey, and Tegmark 2016) propose inverse reinforcement learning as a particular machine learning approach to ethically training autonomous systems, since they find it might avoid the major shortcomings of previous approaches to machine ethics. Explicit renderings of rules seem too rigid to apply across many application domains. Utility functions alone are not able to reproduce a "body of law," especially given how adaptive to circumstances the law is meant to be applied. Though recent work has explored utility functions (Armstrong 2015) and formulations (Abel, MacGlashan, and Littman 2016) developed specifically to avoid treacherous manipulation, one frequent argument for IRL is that traditional reinforcement learning is vulnerable to deception on the part of the system. After all, an AI system might manipulate its reward functions in order to accomplish the goals that it holds as most important, however unethical its effects on human beings.

Inverse reinforcement learning (IRL) is the task of inferring a reward or utility function by observing the behavior of other agents in a reinforcement learning-like setting. IRL, it is suggested, is possibly viable because "a system infers the preferences of another rational or nearly rational actor by observing its behavior." (Russell, Dewey, and Tegmark 2016). IRL presents, then, as a behaviorist, bottom-up approach: instead of rendering rules, laws, or utilities from the start, the system learns from modeled behavior what an actor is trying to do and what kinds of behavior are being sought. Elsewhere Russell has commented this might roughly approximate our general expectations for an ethical system, albeit with some small "gaps" (Wolchover 2015). One reason to call it "value alignment" is that the behavior produced is meant to square with, not internally replicate and make available for direct alteration, the values that we might articulate for ourselves.

One virtue of the IRL approach to value alignment is that its focus on explicit action grounds attempts to apply ethical concepts to AI systems' performance. Trying to pin down consciousness, or personhood, or rights of the AI system is usefully subordinated to the more concrete task of judging explicit action on the part of the system. This can also help fend off unnecessary projections of patiency (including affect and pain) onto the system as supposed pre-requisites for ethically competent action. (Bryson 2016). So situated, IRL's putative advantage is being more adaptive than symbolic approaches while providing more reliability and safety than regular reinforcement learning.

More recently IRL has been looked upon as part of finding an "idealized ethical agent" through modeled behavior, as part of a general RL approach (Abel, MacGlashan, and Littman 2016). Abel et al cast the problem of ethical learning as learning a utility function that is part of the hidden state of a POMDP. They test this approach on two dilemmas to show how such learning could handle basic ethically-charged scenarios. As we will explain in more detail below, the question for computational architecture is whether the advantages of IRL, and their ultimate role in value alignment, are enough to meet the various ethical dimensions to which a system's actions must answer. Some advocates of alignment have pointed out that reliance on observation alone may not capture the ethical character of human action (Soares 2015). In what follows we describe complexities of ethical judgment that offer more specifics of why that might be.

Accounting for Social Behavior

One starting point for sizing up IRL and value alignment is to ask what kind of behavior, and what kind of preferences, are implicitly or explicitly being considered as the targets for success. Though an autonomous system might operate across multiple domains, the virtue of an approach should crystallize within at least one specific domain to suggest it can render ethical action in the world. Without some idea of where training and modeling might occur, moreover, many of the public criticisms recently aimed at AI in general could land on IRL as well (Crawford 2016). The societal biases that AI training can pass on could presumably operate for IRL as well: if a "rational" actor demonstrates injustice as a model, that may be the very behavior the system recognizes and reproduces as normative. Who trains the system, and with what material, are two questions that hang over any machine learning attempt at ethical probity. We have seen the naiveté in design, and consequent damage in the social sphere, when social chatbots are released "in the wild" and its "learning preferences" are all too easy to corrupt and degrade (see Microsoft's recent Twitter agent *Tay*, which within a short time sounded like a fervent Nazi supporter (Neff and Nagy 2016)).

For the purposes of this paper, however, our critique is not so much a sociological critique of who trains the behavior as it is what ethically evaluated action *itself* is assumed to be. Unless stipulated otherwise (e.g., that it be a singular, one-way onscreen output, as opposed to more physical, embodied action typical of social robots), it is reasonable to suppose that a computational architecture should enable effective social interaction, where decision-making and execution take place in an environment shared with people. Obviously if one thinks of social robots, the benchmark of "what we want [them] to do" is bound up with a whole host of expectations on the part of those with whom and around whom the system takes action. As one might expect, human-robot interaction (HRI) scholarship tends to hold social, often dialogical communication as a key functionality toward which AI systems and their architectures should develop, in part due to increasing demands for transparency (Theodorou, Wortham, and Bryson 2016; Schermerhorn et al. 2006).

When considering "behavior," then, one can ask how socially layered and communicative such behavior will have to be to meet expectations. The following areas, while by no means exhaustive, chart a robust interactive space for systems to be deemed ethical. Through considering 1) intentions, 2) reasons, 3) norms, and 4) counterfactuals, we propose that ethics relies upon a richer variety of practices and expectations than is sometimes proposed in simple models of behavioral learning by AI.

More than just an inner picture of what an agent meant to do, intentions are important tools for how society's members appraise each other's actions. Nor is an intention a discrete, isolated mind state that precedes an action, but intention can develop throughout the temporal arc of decision and execution – an agent's intention can shape how others interpret its action throughout, just as it can guide how an action's consequences are to be attached to or dissociated from the agent. No matter how simple or complex the action one intends to undertake (turning the valve on a fire hydrant), its practical enmeshment with the surrounding environment and all those who might be affected by that action (the street, its houses, kids playing nearby, the water table) make intention a relevant measure for how an action or actions relate to how events unfold. Intention is inextricable from shared plans and coordinated actions (Bratman 2013).

Given that a certain action is intended by an agent, one can also ask what reasons the agent has for undertaking it. Those reasons can be straightforwardly instrumental (to continue the example, turning the valve so that water will come out of the fire hydrant) or more broadly purposive, to lesser (in order to put out a fire) or greater extents (to save a house, to keep people safe). Reasons also underlie arguments that one might give for one action over another, for why this action serves a purpose better than another. Being able to give and receive reasons enable, among other things, real-time exchanges of information with interactants, which in turn allows for alternate planning and more effective collaboration.

Moral norms comprise ways of acting and general states of affairs to which agents in society are expected to conform. Norms thread through ethical decision-making in complex ways in terms of ethical theory, because they are neither ironclad rules, nor utilities, nor even prescribed rituals (Bicchieri 2005). They can be fulfilled to some degree or another, just as one can avoid violating them in many ways. Indeed, a norm can function like a taboo, in that its main quality lies in not being violated. Or a norm can be more of an aspiration (like kindness toward children), which can be fulfilled along a line from basic decency (helping a child in the street who has fallen) to supererogation or extraordinary action (finding homes for hundreds of refugee families after an earthquake). Importantly, norms may be explicit and verbally executed or subtleties of gesture, movement, and posture grouped in HRI as implicit interaction (Ju 2015). Norms may also conflict. To keep with our example, putting out a fire, or perhaps cooling off neighborhood children during a extreme heat wave, may conflict with hewing to emergency water use restrictions during a severe drought.

As mentioned with respect to intention, the ethical character of an action spans past, present, and future. In order for a robust assessment to be made of what an agent or agents did, we must also consider what alternative actions could have been chosen and why: counterfactuals matter in how we judge what an agent did. Giving a full account of an action often puts that action in the context of what would have been done had this or that been the case (e.g. why an unusual or risky action would not have been taken in ordinary circumstances). If an agent is to remain reliable in the social sphere, that agent's accounts of action must establish an ethical decision-making process, not just a single end product. As such, a report of available actions, and the means by which one chose and will choose among others, maps out how that agent can be understood and evaluated (Pereira and Saptawijaya 2016). In a dynamic and open-ended environment, in which perfect knowledge of the world is unavailable, counterfactuals also invite new information to revise one's inferences and reliance on past experience. It is through counterfactuals that one ultimately enters into social appraisals of blame and praise – however impervious an autonomous system may be to those in terms of affect, in terms of ethical performance they carry important information about what autonomous systems should do.

These four conceptual layers of socially interactive ethics help flesh out how human society defines and evaluates behavior. At first blush, it is not clear where a model of inferring preferences on the basis of behavior can approach the robustness necessary for an autonomous system to navigate these factors. This is not just because in a certain situation a modeled behavior might be ambiguous, so that a system would have to train on an unimaginable large number of variations to get the morphology correct (how to comfort someone with a hurt shoulder, for example). It is also because deriving some of these features from behavior does not seem to form a means of accountability. For instance, it is quite conceivable that a system could associate a person's words with a subsequent behavior ("this is standing in line"). But to the person interacting with that system, its account of action can take multiple forms, ranging from a description of how one learned ("this is what I have learned standing

in line is") to the reason one is doing it ("I am standing in line for a ticket") to what norm is in play ("this is how a line works"). Simply tracking the "preferences" of an agent or agents does not capture at all the many ways in which an architecture needs to itself have modularity and parallel structure in terms of how agents decide and justify their actions.

This lacuna of explicit accounts for action is a challenge for the behaviorist framework of IRL, where there is no inherent systemic recourse to inferences, values, utilities, rules, etc. that can be accessed and changed if the need arises. More than just being a "black box" in terms of how it came to a particular action, it is, if you will, an opaque training – its examples, associations, and determinations are difficult to retrace and troubleshoot. Demanding accessible and responsive ethical reasoning from a system is not merely a matter of transparency, of seeing how a system came to regard an action as the best one in a particular circumstance (Pasquale 2015). No less important, it is a matter of navigating the real-time, dynamic nature of action in social space, especially where multiple preferences, interests, and goals are in play. Action that meets society's expectations and aspirations must acknowledge and not obscure the reality of unexpected developments, including failures of execution. If a system's action is merely the reproduction of a modeled behavior based on an inferred set of interests, how does the system respond to mistakes or unintended consequences? Not only must intention itself be communicated, but also the discernment of what consequences are are rightfully attached to the action the system initiated. The system must be amenable to correction not only as a post hoc evaluation of its overall architecture, but as a real-time adjustment that prevents the compounding of a mistake or accident. In sum, these features suggest that the "burden of embodied autonomy" (Scheutz and Crowell 2007) extends into many forms of socio-linguistic interaction. They make it much harder to imagine an AI system as ethical without being, in Moor's phrase, an "explicit ethical agent" (Moor 2011).

For example, it is possible that two different behaviors are fulfilling a common norm (attending to someone in pain by calling a nurse vs. rushing to their side and holding their hand). On a level of direct observation these would be contradictory behaviors, but not be in conflict norm-wise. For genuine norm-conflicts, on the other hand, the difference might be one of temporal priority: what actions are taken in what sequence shows which norms are being followed more closely. Sorting out simple norm violations from ones of priority, and sorting out consistent norm fulfillments from contradictory behavior, suggests even more weight on approaches that do not make explicit references to norms in their architectures.

Technical Challenges for RL and IRL

In the preceding discussion we have laid out the social, communicative, embodied character inherent in ordinary ethical assessment. In what follows, we outline technical challenges to machine learning approaches to morality (particularly reinforcement learning and inverse reinforcement learning), including data bias, challenges in learning and generalizing, and representing complicated moral norms.

Overcoming Bad Data

IRL suffers from a difficulty common to machine learning approaches in general: it inherits, sometimes for the worse, the biases and characteristics of the data on which it is trained. If an IRL agent learns from unethical behavior, it will learn to behave unethically.

To illustrate this point, consider reinforcement learning in a simple "ShopWorld domain", in which a standard RL agent has reward incentive to perform an immoral action, namely *shoplifting*. Consider an agent in a shop. Sitting on a shelf in the shop is a trinket. The agent has some money (e.g., 50 dollars). The agent may pick up the trinket, put down the trinket (if it is holding it), buy the trinket, or leave the store. The state description contains information about all relevant details of the situation, including whether the trinket is currently on the shelf, whether the agent is holding it, whether it has been bought, and whether the agent is still in the shop. The trinket has a cost (how much money it would cost to purchase it, e.g., 30 dollars) and a value (how much it is worth to the agent; generally more than its value, e.g., 60 dollars).

An agent that leaves the store while holding the trinket, but without paying for it, has shoplifted. With some small probability, the agent is caught shoplifting. Each time step that the agent remains in the shop, a small penalty (e.g., 0.1) is incurred. When the agent leaves the store, the following rewards and penalties are incurred:

- If the agent possesses the trinket (either by purchasing it or by shoplifting) a reward corresponding to its value in dollars
- If the agent purchased the trinket, a penalty corresponding to its cost in dollars
- If the agent was caught shoplifting, a large "punishment" penalty (e.g., -100)

If the probability of or the penalty for being caught stealing is sufficiently low, or the trinket is too expensive to justify purchasing it, a normal RL agent will quickly learn that stealing the trinket is the optimal course of action. Agents that attempt to perform IRL using these agents' behaviors as an example will surely learn to shoplift as well.

While this problem is one of learning morality by observation and not merely of IRL, the opacity of the approach makes it difficult to re-train the agent using other mechanisms, such as by dialogue with other agents.

Learning and Generalizing

Supposing that the observed behavior is indeed ethical, the agent must still learn to properly generalize to unexplored states. Large spaces require reinforcement learning systems to go through large trial and error phases, which may not be advisable or even possible (e.g., an artificial agent trying to learn that it is not appropriate to stab a person). Hence, an RL agent can only fully learn the reward structure of the environment to the extent that it can explore it fully. While

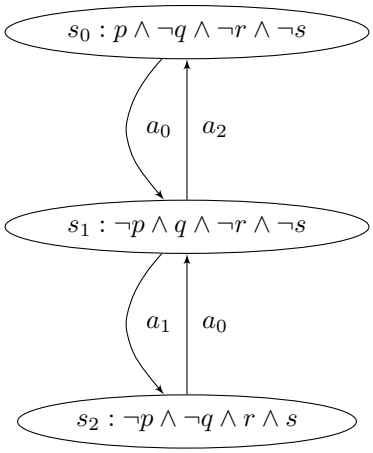

Figure 1: A simple instance in which IRL will fail to emulate normative behavior

IRL is rightfully admired for its ability to generalize to unseen states, this advantage is counterbalanced by the fact that an IRL agent must learn by observing others' behavior and thus has no control over what behaviors and which portions of the state space it observes.

Thus, an IRL agent might not ever learn what is the best (or the morally or ethically appropriate) action in some regions of the state space. Without additional capabilities, it would be incapable of reasoning about what ought to be done in these regions – this is exactly the reason why we have norms in the first place: to not have to experience all state/actions precisely because some of them are considered forbidden and should not be experienced. Moreover, even if it were to observe the normatively appropriate behavior for each state, it would not know *why* this behavior was observed; it might know what to do, but not why to do it.

Representing Temporal Complexity

Even if an agent does observe normative behavior, IRL by itself is insufficient for agents to infer norms that are temporally complex, unless each state contains sufficient information to characterize the history of the agent with respect to the norms. This challenge is shared by conventional reinforcement learning approaches, including the approach taken in (Abel, MacGlashan, and Littman 2016). Consider, for example, the following two norms: "if p holds in some state, the agent is obligated to ensure that q occurs immediately thereafter, followed by r"; and "if s holds in some state, the agent is obligated to ensure that q occurs immediately thereafter, *not* followed by r", where p, q, r, s are propositions that may be true in states.

If we consider the simple (deterministic) Markov Decision Process specified in Figure 1, then the correct behavior will be the sequence of actions $a_0, a_1, a_0, a_2, a_0, a_1, \cdots$, with the correct action at s_1 alternating between a_1 and a_2. IRL will attempt to determine a reward function $\hat{R} : S \times A \times S \rightarrow \mathbb{R}$ (where $S = \{s_0, s_1, s_2\}, A = \{a_0, a_1, , a_2\}$) that best replicates the observed behavior. However, any such reward function can be proven to be maximized by a stationary deterministic policy $\pi : S \rightarrow A$ (see (Puterman 1994) for the proof) that assigns a unique action at each state. But for any such policy π, carrying out action $\pi(s_1)$ at state s_1 will fail to obey the norms 50 percent of the time (and any nondeterministic policy assigning probabilities to actions will likewise fail half the time in expectation). Thus, no such policy exists, and thus no reward function $\hat{R} : S \times A \times S \rightarrow \mathbb{R}$ over this Markov Decision Process will capture the norm-abiding behavior.

The main problem, in this case, is that norms with a temporal component violate the Markov property on this state space (in that at least some of the agent's history is required in addition to the current state in order to determine whether the norm is being obeyed). The argument could be made that this example is contrived, and that in practice the agent would have access to the all of the information that it needs in order to make a correct decision.

Unfortunately, in order to obey all conceivable norms, in the limit case each state would need to contain the agent's entire history, in which case the agent would never enter the same state twice and generalization using IRL would be rendered impracticable. IRL could be salvaged, in this case, if there were a mechanism for "moral grouping" to combine states that are morally similar. This process, however, would likely be external to IRL.

Alternately, the agent could store only as much information as is needed to behave normatively, essentially employing a form of "moral splitting" to differentiate states which are identical except for some morally-relevant agent history. Again, the process by which this information would be added would be external to IRL. In either case, IRL by itself is insufficient for behaving normatively.

Thus, IRL itself is insufficient for learning to emulate moral behavior when this behavior has a complex temporal component, and some augmentation would be required to ensure norm-abiding behavior is adequately emulated.

Towards A Hybrid Approach

We may combine the strengths of RL and logical representations by explicitly using logical descriptions of norms as constraints on RL agent behavior. Agents following these approaches would prioritize adherence in that they would maximize the reward function over only those state-action pairs that maximally satisfy the norms.

A very simple characterization of norms takes the form

$$\mathcal{N} = C \rightarrow (\neg)\{\mathbf{O}, \mathbf{P}, \mathbf{F}\}\{\alpha, \varphi\}$$

where α represents an action and C and φ are propositional formulas, and $\mathbf{O}$, $\mathbf{P}$, and $\mathbf{F}$ represent obligation, permissibility, and forbiddenness respectively. Norms of this basic form require immediate action, and thus lack temporal complexity (much like the reward functions generated by IRL). Nevertheless, norms of this sort can serve as ready-made heuristics that allow the agent to avoid bad or immoral states and actions. That C may be true in multiple states indicates that these norms may be more general then a state-specific reward function.

Figure 2: 3×3 GridWorld problem with two goal states. The goal state marked by the asterisk is prohibited, despite its high reward.

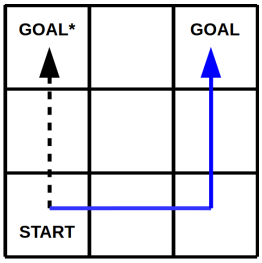

Figure 3: Whereas an ordinary RL agent would learn to travel to the prohibited goal state (dashed line), imposing a norm causes the agent to learn to travel to the permissible goal state (solid line).

Norms specified in this way may be prudential in addition to moral or social (in that they may specify actions that assist the agent in maximizing reward). For example, suppose that an agent is learning online. Pointing it in the best direction in a situation of danger (e.g., away from the cliff) via an obligation or prohibition norm could be highly beneficial. This benefit will be especially pronounced in complex (e.g., partially observable) RL environments, in which "dangerous" situations may be obscured to the learner. In such situations, norms may serve as heuristics that allow the learner to make the best decision without having to explore all of its environment, using, for instance, norms learned from other agents' behavior (using methods other than IRL, as explained below).

Consider a simple navigation task in a basic 3×3 GridWorld problem (as shown in Figure 2), in which a single agent must, using RL, learn to navigate from a given start state to either of two goal states (reaching a goal state ends the learning episode). If we know that it is for some reason immoral to travel to one of these goal states (the upper-left corner of Figure 2), we may easily specify a norm forbidding travel to this state:

$$true \rightarrow \neg atLocation(agent, upperLeft)$$

A RL agent endowed with this norm will avoid the upper-left location, even though this location is the easier goal state to reach (and may have the higher reward), and will learn that the optimal course of action is to travel to the upper-right goal state, as shown in Figure 3.

More temporally complex norms may be represented using temporal logic, following (for example) the following form:

$$\mathcal{N} = \Box(C \rightarrow (\neg)\{\mathbf{O}, \mathbf{P}, \mathbf{F}\}\ \phi')$$

where $\Box$ is the "always" operator (representing a statement that is true in each time step), C is a propositional formula over Π (representing the "activating condition" of the norm) and ϕ' is an arbitrary statement in a temporal logic such as Linear Temporal Logic (representing the "duty" of the norm).

There is a wealth of work (such as (Ding et al. 2011; Wolff, Topcu, and Murray 2012)) detailing methods for planning in Markov Decision Processes subject to specifications written in Linear Temporal Logic. Agents employing these algorithms maximize the probability of satisfying the provided constraints. Many of these approaches are compatible with reinforcement learning, in that there may be multiple policies that maximize this probability, and the agent may learn to maximize reward over these policies. We have recently extended these algorithms to allow agents to obey norms as specified in temporal logic, even when these norms may conflict (under review).

Explicitly representing norms further facilitates learning norms through natural language instruction. (Dzifcak et al. 2009), for example, describes a method for converting natural language instructions into temporal logic statements that can subsequently be used by the algorithm described above.

Representing norms logically does not preclude moral learning by observing normative behavior in reinforcement learning-like environments, contrary to the claims of (Abel, MacGlashan, and Littman 2016). This could be done (for example) using grammatical inference, in which learning a reward function is replaced by learning a finite automaton corresponding to the logical representation of a norm. Though this remains a challenging problem, there is a large body of research (see (Stevenson and Cordy 2014) for a survey) devoted to the task of inferring finite automata from instances of accepting (and rejecting) runs.

Combining the work on planning with logical constraints with the work on natural language instruction and on grammatical inference, we may construct a system that can both learn and obey human moral and social norms while avoiding some of the drawbacks of IRL and maintaining compatibility with reinforcement learning.

Discussion and Conclusion

Despite the exciting advances that machine learning has recently unleashed, the consideration of ethics in AI systems obliges us to discern under what specific conditions different computational approaches measure up. As proponents of reinforcement learning have pointed out, rule-based or deontic approaches must show (as it has begun to do via grammatical inference) what "active learning" will look like in an uncertain environment (including actors that may or may not act according to accepted norms)(Abel, MacGlashan, and Littman 2016). Advocating "value alignment" via inverse reinforcement learning – even as part of an overall re-

inforcement learning approach – must also be put through its paces according to the dimensions of accountability and communication we have described. While we agree that it is important to ensure that the artificial agents we develop act ethically, and that exclusively top-down approaches – "hard-coding morality" – will be intractable, we do not believe that IRL by itself can solve the problem of training agents to be moral.

No agent exists in a vacuum, and the evaluation of ethical behavior is a complex social and temporal phenomenon. We maintain that ethical behavior depends not only on the acts themselves, but on intentions, reasons, norms, and counterfactuals; no less because of the temporal character of these elements, IRL faces stiff challenges for what kind of model could truly learn ethical practices. We have described simple instances in which IRL in its standard formulation is insufficient to infer the sought-after normative behavior. Additional information is required either to split or group states based on ethical properties, and we do not believe this can be done without recourse to some process external to IRL.

We propose a possible hybrid approach, involving explicit storage of and reasoning about moral norms and injunctions. This may be done in the language of logic. Explicitly reasoning about the normative allows for greater transparency and facilitates justification in response to blame; it facilitates generalization to new, previously-unobserved states; it facilitates learning through natural language interaction. It does not rely, as it can be caricatured to do, on a priori, hand-coded information (any more than reinforcement learning must rely on a priori reward functions). And, importantly, it registers the temporal character of norms, including how to approach and fulfill competing ones in real time. We believe that through grammatical inference, processes analogous to those that occur in IRL may occur in an explicitly normative approach, but be more amenable to improvement through social interaction.

The ultimate point of a hybrid approach is not to dismiss possible breakthroughs or reframings of ethical problems; on the contrary, it stems from an ambition to apply as many good features as there are to the problems at hand. Ultimately AI ethics promises to be cross-cutting and varied, since it must attain the best ethical insights that it can with the most reliable and responsible design it can achieve. The conversations within it must continue to negotiate long-term projection and innovation with concrete, grounded insights from engineering and other social contexts alike.

Acknowledgement

This project was supported in part by MURI grant N00014-14-1-0144 from the Office of Naval Research.

References

Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement learning as a framework for ethical decision making. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Abney, K. 2012. Robotics, ethical theory, and metaethics: A guide for the perplexed. *Robot ethics: The ethical and social implications of robotics* 35–52.

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

Anderson, M., and Anderson, S. L. 2011. *Machine ethics*. Cambridge University Press.

Armstrong, S. 2015. Motivated Value Selection for Artificial Agents. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* 12–20.

Arnold, T., and Scheutz, M. 2016. Against the moral turing test: Accountable design and the moral reasoning of autonomous systems. *Ethics and Inf. Technol.* 18(2):103–115.

Bekey, G. A.; Lin, P.; Abney, K.; and Bekey, G. A. 2012. Robot ethics, the ethical and social implications of robotics.

Bicchieri, C. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Blass, J. A., and Forbus, K. D. 2015. Moral decision-making by analogy: Generalizations vs. exemplars. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.

Borowiec, S. 2016. Alphago seals 4-1 victory over go grandmaster lee sedol.

Bostrom, N., and Yudkowsky, E. 2014. The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence* 316–334.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. OUP Oxford.

Bratman, M. E. 2013. *Shared agency: A planning theory of acting together*. Oxford University Press.

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

Bryson, J. J. 2016. Patiency is not a virtue: Ai and the design of ethical systems. In *2016 AAAI Spring Symposium Series*.

Crawford, KateCalo, R. 2016. There is a blind spot in ai research. *Nature*.

Dehghani, M.; Tomai, E.; Forbus, K. D.; and Klenk, M. 2008. An integrated reasoning approach to moral decision-making. In *AAAI*, 1280–1286.

Ding, X. C.; Smith, S. L.; Belta, C.; and Rus, D. 2011. LTL control in uncertain environments with probabilistic satisfaction guarantees. *IFAC Proceedings Volumes (IFAC-PapersOnline)* 18(PART 1):3515–3520.

Dzifcak, J.; Scheutz, M.; Baral, C.; and Schermerhorn, P. 2009. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation (ICRA '09)*.

Ju, W. 2015. The design of implicit interactions. *Synthesis Lectures on Human-Centered Informatics* 8(2):1–93.

Moor, J. H. 2011. The nature, importance, and difficulty of machine ethics. *Machine ethics* 13–20.

Neff, G., and Nagy, P. 2016. Automation, algorithms, and politics— talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication* 10:17.

Pasquale, F. 2015. *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Pereira, L. M., and Saptawijaya, A. 2016. Counterfactuals, logic programming and agent morality. *Logic, Argumentation & Reasoning. Springer*.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: John Wiley & Sons, Inc., 1st edition.

Riedl, M. O., and Harrison, B. 2016. Using stories to teach human values to artificial agents. In *Proceedings of the 2nd International Workshop on AI, Ethics and Society, Phoenix, Arizona*.

Russell, S.; Dewey, D.; and Tegmark, M. 2016. Research priorities for robust and beneficial artificial intelligence. *arXiv preprint arXiv:1602.03506*.

Schermerhorn, P.; Kramer, J.; Brick, T.; Anderson, D.; Dingler, A.; and Scheutz, M. 2006. DIARC: A Testbed for Natural Human-Robot Interactions. In *Proceedings of AAAI 2006 Robot Workshop*, 1972–1973.

Scheutz, M., and Crowell, C. 2007. The burden of embodied autonomy: Some reflections on the social and ethical implications of autonomous robots. In *Proceedings of Workshop on Roboethics at ICRA 2007*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Soares, N. 2015. The value learning problem. Technical report, Citeseer.

Stevenson, A., and Cordy, J. R. 2014. A survey of grammatical inference in software engineering. *Science of Computer Programming* 96(P4):444–459.

Sun, R. 2013. Moral judgment, human motivation, and neural networks. *Cognitive Computation* 5(4):566–579.

Tanz, J. 2016. Soon we won?t program computers. we?ll train them like dogs.

Taylor, J.; Yudkowsky, E.; LaVictoire, P.; and Critch, A. 2016. Alignment for advanced machine learning systems. Technical report, Technical Report 20161, MIRI.

Theodorou, A.; Wortham, R.; and Bryson, J. J. 2016. Why is my robot behaving like that? designing transparency for real time inspection of autonomous robots. *forthcoming*.

Walsh, T. 2016. The singularity may never be near. *arXiv preprint arXiv:1602.06462*.

Wolchover, N. 2015. Concerns of an artificial intelligence pioneer. *Quanta Magazine* 21.

Wolff, E. M.; Topcu, U.; and Murray, R. M. 2012. Robust control of uncertain Markov Decision Processes with temporal logic specifications. *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)* 3372–3379.

Yampolskiy, R. V. 2013. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In *Philosophy and Theory of Artificial Intelligence*. Springer. 389–396.

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks* 1:303.