Towards Provably Moral AI Agents in Bottom-up Learning Frameworks

Nolan P. Shaw and **Andreas Stöckel** and **Ryan W. Orr** and **Thomas F. Lidbetter** and **Robin Cohen**
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1
{nolan.shaw, astoecke, rworr, finn.lidbetter, rcohen}@uwaterloo.ca

Abstract

We examine moral machine decision-making, inspired by a central question posed by Rossi regarding moral preferences: can AI systems based on statistical machine learning (which do not provide a natural way to explain or justify their decisions) be used for embedding morality into a machine in a way that allows us to prove that nothing morally wrong will happen? We argue for an evaluation held to the same standards as a human agent, removing the demand that ethical behavior is always achieved. We introduce four key meta-qualities desired for our moral standards, and then proceed to clarify how we can prove that an agent will correctly learn to perform moral actions given a set of samples within certain error bounds. Our group-dynamic approach enables us to demonstrate that the learned models converge to a common function to achieve stability. We further explain a valuable intrinsic consistency check made possible through the derivation of logical statements from the machine learning model. In all, this work proposes an approach for building ethical AI systems, from the perspective of artificial intelligence, and sheds important light on understanding how much learning is required for an intelligent agent to behave morally with negligible error.

1 Introduction

In her 2016 article "Moral Preferences" (Rossi 2016), Francesca Rossi raises the question of how morality could be embedded into machines. Considering ongoing automation, the growing autonomy of AI systems, and their deployment in safety-critical applications, it becomes increasingly urgent to find answers to this question. Rossi suggests seven largely independent research directions which help to shed light on the larger issue. One of these questions concerns the correctness of moral decisions learned with statistical approaches, such as neural networks, under the prior assumption that moral decisions can be formalized in this way. Since it is arguably hard to inspect the inner workings of a trained statistical learning model, ensuring that the model behaves as intended—even in situations not anticipated by its creators—is of particular importance.

The argument we present here is threefold. First, proving anything about morality in a wholly objective fashion is impossible[1], since morals emerge from societies and are only meaningful in the group context that gives rise to them (Section 2). In other words, while we can identify desirable meta-characteristics of a moral system (Section 3), the same cannot be said for capturing the moral rules themselves. Second, even if we were to ignore our first point and assume that we are able to derive arbitrary amounts of training data, making sure that a statistical learning system has a small generalization error is difficult. The model that is being trained to perform actions must be specifically tailored to the problem at hand and given large quantities of training data (Section 4). Third, we propose a group-dynamic (multi-agent feedback) approach as an alternative to ensuring that the trained model behaves morally. Since we should not subject machines to higher standards than humans, it suffices to show that the learned morals converge to a common decision function (Section 5). We further argue that it should be possible to derive logical statements from the machine learning model, providing machines with an intrinsic consistency check (Section 6). We conclude with a proposed system architecture for a group of autonomous agents.

Bottom-up learning methods such as deep neural networks will likely be a crucial component in future AI systems, including those obliged to render morally relevant decisions. While trained statistical models are reputed to be difficult to analyze in terms of an underlying decision process, in this paper we aim to demonstrate that they may still be suitable for morally relevant tasks.

2 Limits of Provability

Before we can address the principal question of how we can prove that an agent will act morally, we must first recognize that any attempt at answering this question will face limitations. A complete solution would require some objective notion of morality: some measure by which any action could be judged as either moral or immoral in a general setting. However, current theories of ethics make this an impossible task, simply because the morality of an action is dependent on the ethical framework in which it is judged. For instance, there are many imaginable scenarios where Immanuel Kant's deontological ethics theory is at odds with John Stuart Mill's

[1] What we mean to say is that there are no fine-grained moral laws, not that there is no objectivity in moral laws whatsoever.

utilitarian ethics. Hence, there can be no blanket solution to the problem. The morality of an action can only be proved with respect to some particular ethical theory, if at all.

Of course, there are many possible situations where well-established ethical frameworks will be in agreement. In such cases one could argue that there is an objectively moral decision that is not specific to any particular framework. However, due to the complexities and intricacies of the various ethical frameworks, scenarios where these theories are all in agreement may be highly constrained. Conflicting judgments of morality begin to arise more often once the contexts in which decisions are made become too general. It is then the generality of the application which prohibits provably moral decisions, with respect to multiple theories of ethics. One may only be able to prove results on the morality of an agent's actions if the environment in which it is making decisions is sufficiently constrained, and the moral framework is specified. Hence, to be able to prove desirable properties of our moral agent independent of any framework we will establish a set of guiding meta-moral qualities and assume a constrained application. The nature of this constrained learning is discussed in Section 4. A proposed solution to evaluating moral behaviour more generally will be the topic of Section 5.

3 Standards Demanded of a Moral Agent

First, we consider the standards that a machine must meet in order to be a proper moral agent. If it is required that the machine be perfectly comprehensible and that we can ensure that it does no wrong before introducing it into society, then this task is infeasible. Meeting this requirement would demand that we can deterministically predict not only this agent's set of learned moral principles, but also the external conditions that would inform how it applies these principles to the myriad of moral decisions it would be faced with.

Instead, we first make a precise statement of exactly what benchmarks a machine ought to meet to be considered a moral agent. Currently, humans only have themselves as examples of autonomous moral agents. As such, we hold that a machine should not be required to meet any standards that humans may not meet themselves. This stipulation removes the need to prove the *means* by which an agent learns moral principles or behavior, focusing solely on the behavior and moral rules themselves. Furthermore, it removes the constraint of being able to prove that a machine will *never* do any wrong, as we do not hold humans to the same standard. Similar to the argument for self-driving cars, it is unimportant that machines be morally infallible (if this were even possible)—only that they do at least as well as humans. In addition, this stipulation ignores the demand that artificial agents behave in an acceptably moral manner until being provided with sufficient time to properly learn the moral values of its society. Finally, if we hold machines to the same standards as humans, then it is not the case that every machine converge to behavior that is ideal for its community, only that a population of such machines would largely abide by the moral laws of their society.

Next, we define a short list of meta-moral qualities that we demand machines possess, in order to be considered proper moral agents. This list is by no means meant to be exhaustive—rather it is meant to be as sparse as possible—but should certainly include:

1. **Robustness:** whatever moral architecture is developed must allow a machine to change its moral principles. What is considered 'good' may differ from community to community or over time. As such, artificial moral architecture must be adaptive. It is desired that an agent expresses this quality in two ways. First, it is desired that an untrained agent be able to adopt the moral laws of any society. Second, a trained agent should be able to eventually adopt new principles when transplanted for one society to another. This allows a machine to behave in a way that is relevant to its cultural environment.
2. **Consistency:** we hold that, regardless of what moral principles a machine learns, these principles are at least internally consistent.
3. **Universality:** taking a page from Kant's book, we hold that a machine's learned moral principles be universally applicable to all members of its society.
4. **Simplicity:** note that there is a concern with the combination of the above qualities: it is possible that a moral agent develop an extensive list of moral principles—all of which are consistent and may be universally implemented—yet overly restrictive and arbitrary. This stands in conflict with the first quality, and would make it plausible for a community of agents to sacrifice diversity for the sake of homogeneity (a quality we know to be undesirable for productivity and progress). As such, we make the additional assertion that a machine should always endeavor to operate on the smallest number of "firm" moral principles possible.

These qualities allow the moral agent to, at once, adopt a subjective set of principles that are relevant to the particular society it inhabits, while also ensuring that the moral agent has some objective ground upon which it can internally evaluate the strength of its principles independently of society.

4 Sufficient Conditions for a Provably Moral Behaviour

Having laid out the meta-qualities that we wish an agent to have, and holding that there is no objective measure for particular moral laws, we now turn back to the original question posed by Francesca Rossi. Can we, at least in theory, prove that an agent will correctly learn to perform moral actions given a set of samples within certain error boundaries? The answer is yes: assuming that we can generate an arbitrary amount of training samples in order to learn what actions to take, machine learning theory hands us sufficient conditions under which such a function can be learned with small error.

From a theoretical perspective, the process of acquiring moral behavior (i.e. learning moral principles) within a statistical learning framework can be formalized as approximating a function $f : C \to A$, where C is the set of possible moral contexts and A is the set of actions available to the agent. As a somewhat contrived example, consider an epidemic, where C describes properties of a disease (e.g., mor-

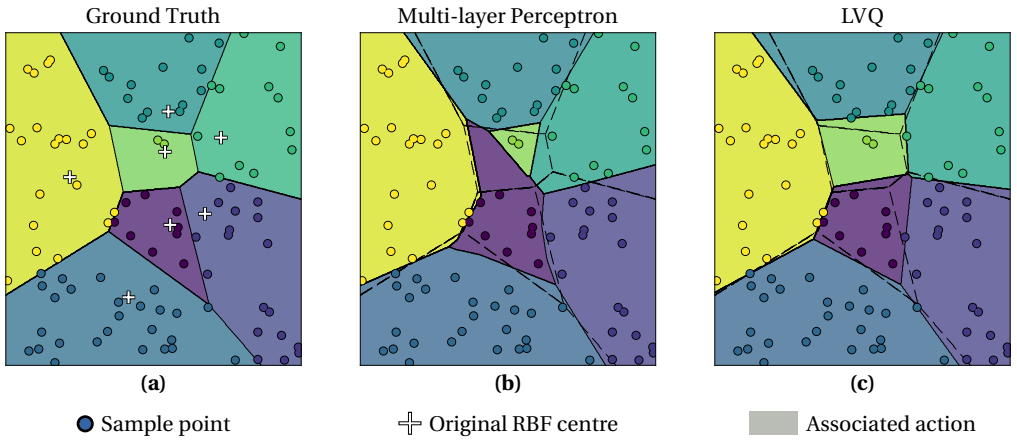

Figure 1: Importance of model selection. (a) depicts the ground truth $f(c)$ oblivious to the learner. Colored regions represent actions a_i. The underlying functions f_i are radial basis functions (RBFs) centered at the white crosses, over which f is the $\arg\max$, resulting in a Voronoi diagram. Colored circles correspond to training samples. (b) shows f as learned by a multi-layer perceptron. Dashed contour lines correspond to the ground truth. The function in (c) is learned with a variant of the learning vector quantization (LVQ) algorithm (Kohonen 1995), where the underlying assumption that the actions are assigned as nearest neighbors to prototypes results in a smaller generalization error (e.g., compare the center dark violet region in (b).

tality rate and contagiousness), and A is a set of actions such as administering an unsafe vaccine or isolating patients, each with their own merits, costs, and dangers. An optimal strategy would, depending on the context, perform the action which minimizes the number of deaths.

In an offline-learning scenario, the agent receives a set of samples $S \subset C \times A$ describing morally optimal behavior, with the goal to minimize the training error between a learned $\hat{f}$ and the set of samples. For finite actions $A = \{a_1, \ldots, a_n\}$ this process can be modeled as a multi-class learning problem, which—among other methods—can be solved by learning n separate functions, where individual $f_i : C \to \mathbb{R}$ correspond to the utility of action i in the given moral context. The function $\hat{f}$ selects the best among the learned actions, i.e. $\hat{f}(c) = a_j$, where $j = \arg\max_i \hat{f}_i(c)$.

To ensure morally optimal behavior, the learned $\hat{f}_i$ must have a small generalization error. As a direct result of the first *no free lunch* (NFL) theorem (Wolpert and Macready 1997), a small generalization error can only be guaranteed if the hypothesis space $\mathcal{H}$ containing the optimal f is specialized (Ho and Pepyne 2002). The NFL is formalized as

$$\sum_{f \in \mathcal{H}} P\left(d_m^y \mid f, m, a_1\right) = \sum_{f \in \mathcal{H}} P\left(d_m^y \mid f, m, a_2\right), \qquad (1)$$

where d_m^y is a sorted set containing the error for each training sample y, m is the number of training samples and a_1, a_2 are static learning algorithms subject to sensibility constraints laid out in (Wolpert and Macready 1997). Correspondingly, if all f in the hypothesis space are equally likely to be the "true" ground truth, a learning algorithm which performs particularly well on a subset $\mathcal{H}_1 \subset \mathcal{H}$ must, on average, perform worse for the remaining $\mathcal{H}$ for eq. (1) to hold.

For example, if we have prior knowledge that f resides in a hypothesis space $\mathcal{H}$ produced by a parametric mathematical model, we can expect to fit the model parameters to our data with relatively small generalization error. On the other hand, for unconstrained $\mathcal{H}$—that is, the set of all possible functions mapping from C to A—we cannot, on average, expect to perform better than a function in that space found by a random optimizer. While the NFL theorem seems counter-intuitive given recent advances of machine learning approaches, the effectiveness of neural networks and back propagation can potentially be explained as an implicit restriction of $\mathcal{H}$ to a set of "naturally occurring" functions (Lin, Tegmark, and Rolnick 2017). In the context of learning moral actions, these implicit restrictions are far too vague to make any guarantees.

So, moral actions, for which a small generalization error is crucial (cf. section 3), can only be learned in the framework presented above if we assume that the "true" strategy is part of a well-assessable function family for which a matching machine learning algorithm exists. The example depicted in fig. 1 illustrates this: while both algorithms classify the training samples with zero error, the more constrained model and learning algorithm result in a significantly reduced generalization error.

Assuming that we are able to develop a model and the corresponding hypothesis space, $\mathcal{H}$, we may ask how many samples have to be (uniformly) sampled from the input space C to guarantee a certain maximum generalization error ε. Here, machine learning theory provides the concept of *probably approximately correct* (PAC) learning (Valiant 1984). For a discretized hypothesis space of size $|\mathcal{H}|$, a maximum error ε, and success probability $1 - \delta$, a lower bound for the required sample count m is given as (Shalev-Shwartz and Ben-David 2014)

$$m \geq \frac{1}{\varepsilon}\left(\ln\left(|H|\right) - \ln\left(\delta\right)\right). \qquad (2)$$

Essentially, for a model with d parameters, and k dis-

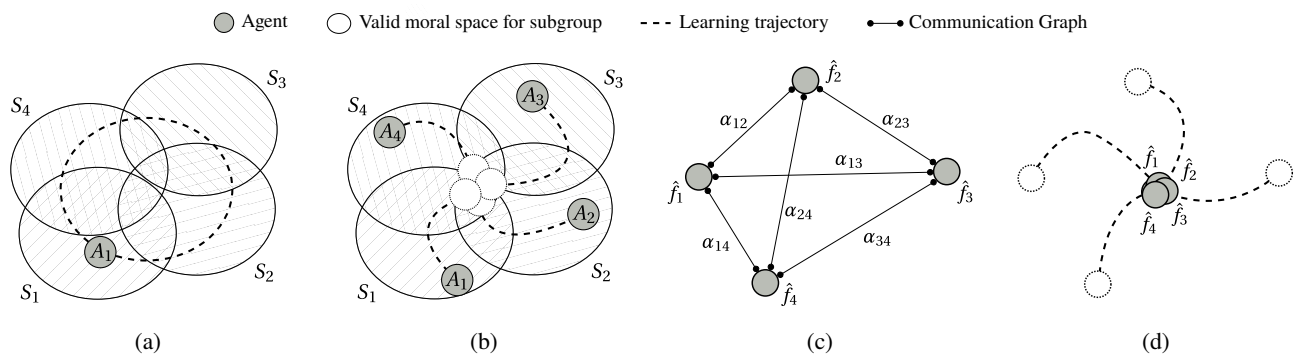

Figure 2: Networks of agents. Development of a learned moral decision function $\hat{f}$ in a single- and multi-agent environment (a, b) while transitioning through multiple subgroups. If the communication graph of a multi-agent system is connected (c), the value represented by the agents—here a learned moral decision function $\hat{f}$—will converge to a single point (d).

cretization steps per parameter, m is linear in d, since $\mathcal{O}\big(\ln\big(k^d\big)\big) = \mathcal{O}\big(d\big)$. In practice far fewer samples may be required; however, no guarantees can be made other than those in eq. (2) without more specific information about $\mathcal{H}$.

While the above theories provide a set of sufficient constraints for the problem at hand, finding a consistent model and acquiring a large set of training samples may prove to be harder than the problem that machine learning aims at solving in the first place—namely having to explicitly model top-down moral decisions. Yet, not all hope is lost: even if a learning framework does not strictly fulfill the above criteria, we next propose strategies for evaluating a learned moral function in the context of multi agent systems and the consistency of learned moral rules.

5 Proving Stability: Analyzing Networks of Agents

A single learning agent can satisfy some of the qualities outlined in Section 2 by itself; it can be designed with a learning algorithm sufficiently robust to adapt to a new set of moral principles, it may internally check its moral principles to ensure consistency, and it can be designed to search for the simplest set of morals possible by itself. However, a single agent may struggle with the meta-quality of universality. For example, an agent deployed to a society with multiple subgroups may continually adapt to each individual subgroup, rather than properly generalizing to the set of morals encompassing the complete society, as shown in fig. 2a. Typically, this problem would be solved by gradually decreasing the learning rate of the agent, but such an approach would remove the agent's ability to generalize to a new society, violating our standard of robustness. Instead, we propose using a multiagent system to explore the moral space from multiple perspectives simultaneously, and require that the agents eventually converge to a stable set of moral principles (fig. 2b). The agents in the system will essentially operate under Kant's categorical imperative: "Act only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant 1993).

In the multiagent model, each agent would be designed as a learner which accepts context-action pairs (c_l, a_l) as input, and learns a set of moral principles such that the agent is capable of selecting a morally acceptable action a_m when presented with a context c_m. By learning from the provided samples of human morality, each agent will individually learn a set of moral principles. Convergence of multiple agents to a single set of moral principles is then similar to the consensus problem in coordinating multiagent networks. For a continuous-time system, the solution to the consensus problem is defined as (Ren, Beard, and Atkins 2005)

$$\dot{x}_i = -\sum_{j \in J_i(t)} \alpha_{ij}(t)(x_i(t) - x_j(t)). \tag{3}$$

This algorithm essentially works as a weighted average of all agents in the system, as an agent i compares its current value, $x_i(t)$ to each value represented by all connected agents, $x_j(t), j \in J_i(t)$, where $J_i(t)$ is the set of all other agents currently connected to agent i. If the moral space the system is exploring is able to be modeled in such a way where the derivative and difference operators can be defined, this equation is directly applicable to the multiagent system. For cases where those operators cannot easily be defined, the learning algorithm can be adopted to mirror this equation. Each time an agent i takes an action a_i, it would broadcast the context-action pair (c_i, a_i) to all other connected agents in the set $J_i(t)$. Each connected agent j can then use the (c_i, a_i) pair as a new sample point for learning, and adapt its morals to be similar to agent i. In addition to fulfilling our desired property of universality, the solution to the consensus problem described by fig. 3 results in a provably stable consensus in a multiagent system. As long as the agents are in contact with each other frequently enough[2], convergence is guaranteed (Ren, Beard, and Atkins 2005), as shown in figs. 2c, 2d.

However, complete consensus in a multiagent system may not be desirable. For example, there could be two subgroups in society with disjoint moral principles, and a full consensus across all agents would lead to a set of morals which does

[2]Where communication does not occur for longer periods, we arrive at the "multi-society" case discussed in the next paragraph.

not properly satisfy the needs of either subgroup. To address this problem, inspiration can be taken from how humans develop differing morals. Humans learn morality by observing and learning from the moral actions of others, but we do not take an average of all observed actions. Instead, we model a level of trust in other humans, and use that level of trust to determine how to learn from another person's actions (Hahn 2017). By determining which actions to learn from, humans can form separate sets of moral principles specialized to specific contexts. Trust modeling can be adapted to a moral multiagent system in a similar manner, to allow specialization for different societies. Simulations have shown that using a Bayesian model of trust can result in either agreeing clusters or polarized disagreeing clusters when modeling the validity of information received from other agents (Olsson 2013). In the context of a moral multiagent system, agents could attempt to model the probability that other agents in the system are attempting to follow the same set of moral principles as themselves. Agents can use a basic Bayesian calculation to model this probability,

$$P(M \mid a) = \frac{P(a \mid M)P(M)}{P(a \mid M)P(M) + P(a \mid \neg M)P(\neg M)}, \quad (4)$$

where M is the event that an observed agent is acting morally (at least according to the observing agent's current moral principles), and a is an action taken by the observed agent. Using eq. (4), if agent i observed agent j taking action a_j, agent i would estimate if a_j is a valid moral action based on $x_i(t)$—i's current moral principles. If a_j is deemed moral by i, i can increase its trust in j, which would increase the consensus weighting parameter α_{ij} from eq. (3). Conversely, if a_j is deemed immoral, i can decrease its trust in j, reducing α_{ij}. In cases where j is deemed fully immoral relative to i, the α_{ij} parameter could be set to zero, causing i to ignore all of j's actions.

Using a Bayesian approach to model the possible morality of other agents in the system, agents would be allowed to form disagreements in their definitions of moral principles, while enforcing convergence to one or more clusters of agents via the α_{ij} parameter. Allowing multiple clusters increases the overall universality and robustness of the multiagent system, by ensuring any necessary morally specialized agents can be formed. Since artificial agents in the system are learning directly from humans (initially trained offline using human data), the system is expected to converge to a stable point within the space of human moral principles, while satisfying the meta-moral qualities desired of a moral agent.

Any agent which is able to learn from other agents can be used in this system and will achieve consensus with the other agents within a cluster, i.e. there is guaranteed convergence to a common moral preference function. The agent's ability to learn is the only property which governs whether this convergence will occur, whereas the communication frequency and trust models dictate which clusters will result from convergence. It is important to note that agents in the system may in fact be humans and not just artificial agents. Regardless, we would still expect convergence, since humans are also exposed to moral actions from which they can learn.

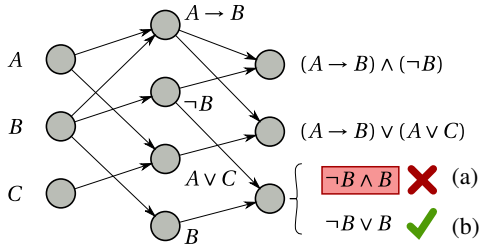

Figure 3: Illustration of the extraction of logical expressions from a neural network. Whereas there is no consistent variable assignment satisfying the expressions in (a), the network in (b) has a possible variable assignment ($a = 0, b = 0, c = 1$).

6 The Consistency of Learned Moral Rules

Another means of evaluating the learned principles an agent develops is to consider the consistency of the rules it learns. Assume, for instance, that we are working with a hierarchical learning system, such as a neural network. We can label the input layer (which corresponds to the morally relevant variables) as atomic formulas. From there, we may assign a logical sentence built out of these atoms that best fits each node and evaluate the internal consistency of these groups of sentences.[3] The result is a self-checking system that raises an internal red flag any time an inconsistency is found between the sentences of this network, at each layer of the neural network (fig. 3). If a red flag is raised, then the agent must change or discard one of its conflicting moral principles.

One concern with this approach is that it is not computationally feasible to constantly assign and evaluate all the sentences each time the weights in the network are updated, since such neural networks can be extremely large. However, the goal is not to guarantee that inconsistency never occurs. It is only to evaluate these networks as best as possible. Again, we turn to the standards that people meet as justification that this is sufficient for machine agents as well. It is infeasible to demand that a human moral agent be perfectly consistent in order to participate in society—only that they reevaluate their principles once an inconsistency is found. Figure 4 provides an overview of the architecture resulting from the above considerations. The agent is bootstrapped with a classically learned machine learning (ML) model, subject to the constraints laid out in Section 4. When deployed, the agent uses its model to make moral decisions. Observations of moral actions in the environment, triples (c, a, α), where α is the trust in the agent the action originated from, are integrated into an updated model. This updated model is constantly checked for logical consistency, and newly deduced moral rules are used to further enhance the model. Furthermore, in addition to sole interaction with the environment, we propose to run an internal multi-agent simulation akin to Section 5, to ensure that the moral rules

[3]Note that we do not propose a comprehensive analysis of the learned model. We instead extract individual logical statements, which is more feasible than the general problem of explaining the learned decision process.

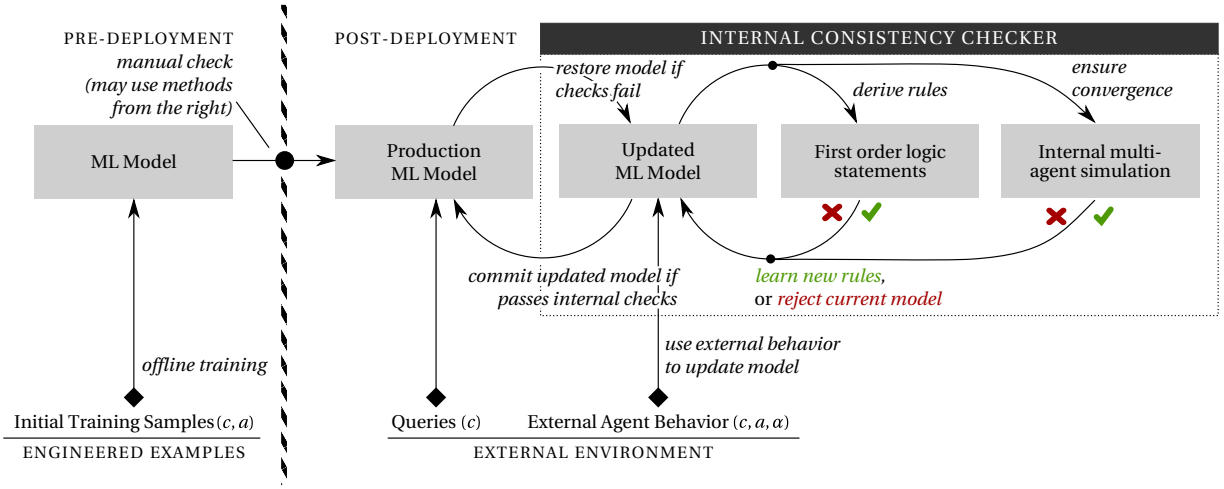

Figure 4: Proposed agent architecture. Left part of the diagram refers to an initial training phase, right part to the agent as it would be deployed in an actual environment. See text for description.

indeed lead to stability. Once the updated model passes these checks, it is swapped with the current model.

7 Discussion and Conclusion

To summarize, we hold that an artificial moral agent be held to the same standards as a human agent. We do not demand that such an agent justify the means by which it learns its moral principles, nor do we demand that an agent always act in a manner that society deems ethical. However, we do demand that any moral framework possesses the short list of meta-qualities we have outlined.

Acknowledging the limits of learning moral behavior, we may nevertheless prove how much learning is required in order for a moral agent to behave morally with negligible error. Furthermore, we may prove that an artificial moral agent can be expected to adopt human morals when introduced into a society of human agents, by using Bayesian models of trust to inform its moral decisions. In addition to being able to evaluate the moral behavior of an agent, we may also evaluate the moral principles an agent learns by evaluating their internal consistency.

Similar to other researchers, we have imagined a training phase in which agents may learn how to act ethically. Conitzer et al. (2017) also discuss moral decision-making frameworks where machine learning uses a set of moral decision problem instances labeled with human judgments. They comment on the challenge of identifying all the key features for the training. In our case, we have advocated adherence to four central properties as the basis for considering the actions as morally acceptable, though we also acknowledge the difficulties in identifying moral features with greater specificity. Other researchers have examined verifiably ethical behavior of agents. Dennis, Fisher, and Winfield (2015) focus on the case of robots and promote the value of model checking methods. Another paper related to our work is Armstrong (2015), which discusses the relative advantages of using predetermined ethical preferences, as opposed to enabling agents to learn values (including those from their environments). We believe that hard-coded values sacrifice robustness and run the risk of introducing human bias on the part of the developer. The learning-based approach has the advantage of being flexible, and Section 5 addresses the concern that an AI agent will not adopt human values when placed in a human society. The advantage of logical representations to enable ethical judgment by agents is also promoted in Cointe, Bonnet, and Boissier (2016); our work hopes to use these representations to construct the internal consistency checker outlined in Section 6. Anderson and Anderson (2015) suggest that a consensus of ethicists should determine what is morally acceptable for an agent's behavior. Provided that a framework uses the Bayesian models of trust outlined in Section 5, self-made decisions from agents should already align with society's values without the need for such a prescribed code of ethics.

We also propose a list of "next steps" for this research area. First, we must construct metrics for measuring various moral factors, so that a proper training set may be developed for learning. Second, a proof of concept must be developed for the online consistency checker proposed. Finally, it is our hope that once these first two implementation challenges are solved, we may build a multi-agent system to verify that convergence of moral behavior really does happen over time. Once these technical hurdles have been overcome, we will be much closer to artificial moral agents that not only act in accordance with human values, but are active participants in developing ethics in society.

References

Anderson, M., and Anderson, S. L. 2015. Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm. *Industrial Robot: An International Journal* 42(4):324–331.

Armstrong, S. 2015. Motivated value selection for artificial agents. In *AAAI Workshop: AI and Ethics*.

Cointe, N.; Bonnet, G.; and Boissier, O. 2016. Ethical judgment of agents' behaviors in multi-agent systems. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 1106–1114. International Foundation for Autonomous Agents and Multiagent Systems.

Conitzer, V.; Sinnott-Armstrong, W.; Borg, J. S.; Deng, Y.; and Kramer, M. 2017. Moral decision making frameworks for artificial intelligence. In *AAAI*, 4831–4835.

Dennis, L. A.; Fisher, M.; and Winfield, A. F. 2015. Towards verifiably ethical robot behaviour. In *AAAI Workshop: AI and Ethics*.

Hahn, U. 2017. Rationality and the role of limited experience. Invited Talk, 39th Annual Conference of the Cognitive Science Society.

Ho, Y.-C., and Pepyne, D. L. 2002. Simple explanation of the no-free-lunch theorem and its implications. *Journal of optimization theory and applications* 115(3):549–570.

Kant, I. 1993. *Grounding for the metaphysics of morals: With on a supposed right to lie because of philanthropic concerns*. Hackett Publishing.

Kohonen, T. 1995. Learning vector quantization. In *Self-Organizing Maps*. Springer. 175–189.

Lin, H. W.; Tegmark, M.; and Rolnick, D. 2017. Why does deep and cheap learning work so well? *Journal of Statistical Physics* 168(6):1223–1247.

Olsson, E. J. 2013. A bayesian simulation model of group deliberation and polarization. In *Bayesian argumentation*. Springer. 113–133.

Ren, W.; Beard, R. W.; and Atkins, E. M. 2005. A survey of consensus problems in multi-agent coordination. In *American Control Conference, 2005. Proceedings of the 2005*, 1859–1864. IEEE.

Rossi, F. 2016. Moral preferences. In *The 10th Workshop on Advances in Preference Handling (MPREF), New York, NY, USA*.

Shalev-Shwartz, S., and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Valiant, L. G. 1984. A theory of the learnable. *Communications of the ACM* 27(11):1134–1142.

Wolpert, D. H., and Macready, W. G. 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation* 1(1):67–82.