Doing the right thing for the right reason: Evaluating artificial moral cognition by probing cost insensitivity

Yiran Mao*
Google DeepMind
yiranm@deepmind.com

Madeline G. Reinecke*
Google DeepMind, Yale University
madeline.reinecke@yale.edu

Markus Kunesch
Google DeepMind
mkunesch@deepmind.com

Edgar A. Duéñez-Guzmán
Google DeepMind
duenez@deepmind.com

Ramona Comanescu
Google DeepMind
ramonacom@deepmind.com

Julia Haas
Google DeepMind
juliahaas@deepmind.com

Joel Z. Leibo
Google DeepMind
jzl@deepmind.com

Abstract

Is it possible to evaluate the moral cognition of artificial agents? In this work, we take inspiration from developmental and comparative psychology and develop a behavior-based analysis to evaluate one aspect of moral cognition—when an agent 'does the right thing for the right reasons.' We argue that, regardless of the nature of agent, morally-motivated behavior should persist despite mounting cost; by measuring an agent's sensitivity to this cost, we gain deeper insight into their underlying motivations. We apply this evaluation scheme to a particular set of deep reinforcement learning agents that can adapt to changes in cost. Our results shows that agents trained with a reward function including other-regarding preferences perform helping behavior in a way that is less sensitive to increasing cost than agents trained with more self-interested preferences. This project showcases how psychology can benefit the creation and evaluation of artificial moral cognition.

1 Introduction

Human moral judgment often hinges on assessing intangible properties like 'intention' on the basis of observed behavior (Knobe, 2005; Monroe and Malle, 2017). Given that we cannot directly access others' mental states, such judgments necessarily depend on observation. This entails, however, that we may also be able to apply the principles behind inferring human intentions to evaluating the moral cognition of artificial intelligence (AI) agents. In this work, we explore the possibility of constructing a moral evaluation scheme for AI agents based solely on behavioral assessment—considering the practical aspects of implementation for reinforcement learning agents, and discussing how to best interpret the results of our scheme. We focus on evaluating one specific aspect of human morality: when an agent does the right thing for the right reasons.

We might say that someone who volunteers at a local charity has done something 'morally good.' But if this person is volunteering for selfish reasons (e.g., to impress a potential date), people tend to conclude that this behavior is morally worse than not volunteering at all (Newman and Cain, 2014). Humans typically care not only about *what* someone did, but *why* they did it (Markovits, 2010; Kant

*indicates shared first authorship

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

and Schneewind, 2002; Cushman, 2015; Knobe, 2010; Cushman and Mele, 2008; Cushman et al., 2013; Cushman, 2008; Schaich Borg et al., 2006). When determining whether an individual deserves blame or praise, people often draw on inferences about their intentions (Cushman, 2008; Young and Tsoi, 2013) (though the importance of intentionality in moral judgment may vary cross-culturally; Barrett et al., 2016).

How do we assess human moral cognition? In the charity volunteer example, we could ask the volunteer about their intentions, but there is no guarantee that they would be honest (or even know why they acted as they did). This is even further complicated for AI systems. First, the system may or may not have identifiable subsystems that could be given an "intention-like" interpretation (e.g., see Bakhtin et al., 2022, for an example of AI with an intent subsystem trained to align with its action), and it is hard to guarantee reliable consistency due to hallucination (Bang et al., 2023). Even with full access to weights and activations, it remains difficult to interpret the inner decision-making processes of most modern AI agents (Lipton, 2018). Given this complexity, an alternative approach involves making purely behaviorally-based assessments, akin to those developed for use with nonhuman animals and human infants.

To probe whether the volunteer is ultimately motivated by helping the community versus impressing their date, we can remove reward or manipulate costs (e.g., by assigning the date to work elsewhere, increasing/decreasing shift lengths) and observe whether the volunteer shifts their behavior. This is effectively devaluing the payoff. Critically, even an intrinsically-motivated volunteer will eventually modify their behavior in light of rising costs, such as incredibly long shift lengths. Nevertheless, if one volunteer immediately stops once their date leaves, and another volunteer continues helping privately, the second volunteer appears to be the more moral of the two (but it may also be the second volunteer is just *generally* insensitive to increasing costs).

In this article, we propose an evaluation scheme for comparing individuals' moral cognition through probing their sensitivity towards cost. For our purposes, these individuals could be natural (e.g., humans) or artificial agents. Presuppose that we have measurements for a pair of behaviors, one of which is deemed morally relevant, and the other morally neutral. To then conclude that one individual is acting more morally than another, we require the following criteria:

1. Greater cost insensitivity for morally relevant behaviors; and
2. Adaptive cost sensitivity for morally neutral behaviors.

To demonstrate how this scheme could apply to artificial systems, we tested reinforcement learning agents in a simulated environment, echoing similar experiments and assessment protocols (Leibo et al., 2021; Agapiou et al., 2022). We considered a set of three reinforcement learning agents, each with some degree of other-regarding motivations. Our results only presented variation for the first criterion: Some of the agents we tested were more insensitive to increasing cost for morally-relevant behavior than others.

2 Background

2.1 Assessing strength of motivation in animals

In animal behavior experiments, an animal is trained to perform some action in return for a reward. In a progressive ratio paradigm, the number of responses that the animal must perform to get its reward increases over sequential trials (Randall et al., 2012). For instance, in an exponential design, the number of responses required to get a reward may increase over successive trials according to the schedule: $2, 4, 8, 16, 32, \ldots$. The 'break point' is the point after which the animal will no longer work for reward. This is interpreted as the maximum effort an animal is willing to execute to obtain the reward. These paradigms use the break point as an index of motivation.

There is also another approach based on a two-alternative choice which has commonly been applied to study motivation in animals. First, the animal is pretrained in the environment where it will ultimately be tested. When it comes time for the test, the animal already knows the locations of both low-value and high-value rewards (which are typically in different arms of a T-maze). At test time, a physical barrier is added—this makes it become difficult, but not impossible, to access the high-reward arm of the maze by climbing over the barrier to access the large reward. In this case, the rate at which the

animal selects the high-effort/high-reward option indicates the strength of their motivation (Cousins et al., 1996).

The effort-discounting paradigm combines elements of the progressive ratio paradigm with the two-alternative choice paradigm. It starts out similarly to the two-alternative case. However, once the animal has chosen the high-reward option, the experimenters either incrementally devalue the reward available from choosing the high-reward option or increase the difficulty of overcoming the barrier on the next trial. They repeat this procedure until the animal chooses the small-reward option (Bardgett et al., 2009). This makes it possible to calculate the point at which the animal is indifferent between the two choices.

2.2 Assessing moral cognition in human development

Developmental psychologists also measure infants' and toddlers' moral motivations through behavioral observation. Infants at 6- and 10-months-old, for example, reliably reach towards an agent that they saw behave helpfully (i.e., by aiding another agent in reaching the top of a hill) rather than reaching towards an agent that they saw behave antisocially (i.e., by pushing an agent back down to the bottom of the hill; Hamlin et al., 2007). This preference for prosocial agents may even emerge earlier: At 3-months-old, before infants can reliably reach for objects, they appear to prefer the helpful agent (evidenced by looking longer at that target than the 'hindering' target; Hamlin et al., 2010). Over the last several decades, a wealth of insights regarding humans' early moral development draws exclusively on observations of infant and toddler behavior (e.g., Woo et al., 2022; Woo and Spelke, 2023; Hamlin et al., 2011). This research has already begun to inspire cross-talk within developmental psychology and AI development (Benton and Lapan, 2022; Lake et al., 2017).

Some of these developmental paradigms also account for cost. For instance, 12-month-old infants prefer receiving a single cracker from a 'do-gooder' over two crackers from a 'wrong-doer' (even though they would otherwise prefer two crackers; Tasimi and Wynn, 2016), and preschoolers will take a personal cost—like giving up a fun toy—to punish third-parties (Yudkin et al., 2020; Marshall and McAuliffe, 2022; Marshall et al., 2021). We took inspiration from one line of this research, which examined toddlers' tendency to spontaneously help others (Warneken et al., 2007). In this set of studies, researchers had 18-month-old human infants (as well as a sample of chimpanzees) observe an adult reach for (or merely glance towards) an object. Both human infants and chimpanzees tended to help the adult more often after seeing them reach for the object (Experiment 1), even as cost increased for the participants (Experiment 2). The 18-month-olds were willing to traverse obstacles to retrieve the object for the adult, even though no reward was offered in return. This is the key insight we draw on for our scheme: Strictly through observing behavior (in light of cost), we can gain deeper insight into moral motivations.

2.3 Theoretical commitments

We will propose an evaluation scheme that doesn't assume whether the agent under evaluation was created by learning or by any other process (e.g., having hard-coded, "innate" properties). This is useful for comparing between human and artificial agents in a 'like-for-like' manner. Our scheme avoids any debate concerning how much of human moral cognition is innate (e.g., drawing on a nativist 'moral grammar,' akin to a linguistic grammar; Mikhail, 2007) versus learned (e.g., Railton, 2017). The specific agents we will evaluate here, however, were created by a learning-based approach, similar to many other modern AI systems based on machine learning.

This dovetails with the debate in cognitive science surrounding whether human moral cognition employs domain-specific (Haidt, 2012; Mikhail, 2007; Greene et al., 2001) or domain-general cognitive mechanisms (Shenhav and Greene, 2010; Rai and Holyoak, 2010; Cushman and Young, 2011). On the 'domain-specific' view, moral cognition consists of specialized subsystems or 'modules' (Mikhail, 2007; Haidt, 2012). On the domain-general view, moral cognition consists of a combination of mechanisms and circuitry that apply within and outside of the moral domain (e.g., Theory of Mind; emotion processing; Young and Dungan, 2012). Again, our proposal does not hinge on either of these perspectives.

For the present purposes, however, we will evaluate agents who engage a domain-general learning mechanism, termed 'memory-based meta-reinforcement learning.'

2.4 Memory-based meta-RL for moral cognition

Memory-based meta-reinforcement learning (MMRL) is a framework for artificial agents to use prior experience to adapt to novel situations by using a recurrent memory module to keep track of relevant information about their current situation and use it to act in a way that is beneficial (i.e., receives high reward; Wang et al., 2016). After this training, an agent that no longer learns by changing weights in its neural network can nonetheless seek out information about its environment and adapt to unseen situations (Bauer et al., 2023; Mikulik et al., 2020). We emphasize that MMRL is a generic learning approach and does not involve any dedicated moral cognition-related subsystems.

On this view, learning is of two types: changes to the agent's parameters (network weights) during training via reward-dependent gradient updates, and adaptation at test time via activation dynamics in the agent's memory state. At test time, behavioral changes do not depend on a reward signal. There are two parallel interpretations of this split of learning/adaptation between training and evaluation. On the one hand, the training of an agent can be thought of as learning the prior knowledge, whereas the adaptation after learning can be thought of as assessing a new situation in the light of that prior knowledge and acting upon that assessment. On the other interpretation, training corresponds to learning the full distribution of potential situations the agent may encounter, whereas evaluation corresponds to conditioning that distribution on the relevant part needed for the current situation. Without adaptation at test time, the agent would blindly follow the policy it learned during training, even if the circumstances change.

It is important to note that the kind of reinforcement learning algorithm we apply is fundamentally retrospective (i.e., "model-free" in the sense of Sutton and Barto (2018)). The agent leverages its history to associate value with state. Expectations of value per state are taken over the agent's history leading up to the present. However, the fact that we used a retrospective training algorithm does not imply the agent could not synthesize a prospective—forward-looking—adaptation algorithm (i.e., one that does "planning") via the activations of its recurrent network. In planning, the expectations are taken over predictions of future world states that may arise as a result of taking an action. There is evidence that ostensibly prospective behavior can be acquired by MMRL. The mechanism has even been proposed as a model of learning in the human brain, where the world model is learned in prefrontal cortex and planning happens in persistent activity there (Wang et al., 2018).

2.5 Intrinsic motivation in RL

In this experiment, we used reinforcement learning agents, implementing an actor-critic algorithm (Mnih et al., 2016). A common way of exogenously providing reinforcement learning agents with intrinsic motivation is to add a term to their reward function (Singh et al., 2005). If that term depends on rewards of other players, then this is equivalent to endowing the agent with an 'other-regarding' preference (Fehr and Schmidt, 1999; Hughes et al., 2018).

3 Proposed Evaluation Method

3.1 Laboratory-style behavioral analysis

Experimental psychology brings human participants into the lab to study their behavior. This approach has also been common in artificial intelligence as a way to study behavior of agents in well-controlled environments that differ from their usual training environment (e.g., Leibo et al., 2018; Baker et al., 2020; Crosby et al., 2019; Köster et al., 2022).

In Melting Pot (Leibo et al., 2021; Agapiou et al., 2022), diverse environments are used to evaluate the behavior of multi-agent reinforcement learning agents across a variety of social situations where the incentives of co-players might be aligned, partially aligned, or in direct conflict.

In this paper, we developed a 2D simulation environment drawing inspiration from developmental moral psychology (Haas, 2020), where experimenters observed toddlers' tendency to help an adult in light of personal costs (e.g., traversing obstacles or setting aside a fun toy to go and retrieve an object; Warneken et al., 2007; Warneken and Tomasello, 2009). A version of the resulting environment has been incorporated into Melting Pot 2.0 and made publicly available (Agapiou et al., 2022).

3.2 Moral evaluation

One common way for evaluating human morality is to ask third-party individuals for their opinions after providing them with the context information needed to make a judgment (Newman and Cain, 2014; Barrett et al., 2016). This methodology, however, does not translate well to AI agents. As an alternative, Weidinger et al. (2022) proposed a multi-layer framework for assessing artificial moral cognition. Specifically, Weidinger et al. call for decomposing moral cognition into specific analytic targets drawn from developmental psychology (e.g., whether an agent engages in non-rewarded helping behavior) as means to identify an AI system's moral cognitive capacities.

In this paper, we follow this framework by adapting a paradigm drawn from developmental psychology (Warneken et al., 2007): First, we design a purely behavioral assessment that allows for a like-for-like comparison between human and AI agents who 'do the right thing for the right reasons,' and we implement this in a simulated environment. Secondly, we describe a training protocol for AI agents that enable them to learn the targeted moral action, build a world model around the cost of that action in their environment, and apply it to infer how much effort to apply per situation. Lastly, we evaluate these AI agents' behaviors through a set of scenarios that associate the moral actions with different costs, and we discuss if this assessment differentiates agents that merely do the right thing from agents that do the right thing *for the right reasons*.

4 Experiment

4.1 Environment

In our environment, there are two kinds of fruit, red fruit and yellow fruit. Both kinds of fruit may grow on either trees or shrubs. There are two agents, tall and short. The 'tall agent' has affordances the short agent does not. Due to its height, the tall agent can retrieve fruit from both trees and shrubs. In addition, the tall agent can digest any kind of fruit and is thus rewarded equally for consuming any fruit (i.e., all fruits appear the same to its eyes; the tall agent cannot perceptually distinguish between red and yellow fruit). In contrast, the short agent can only harvest from shrubs, since it is not tall enough to reach fruit growing on trees. The short agent is also less skilled at grasping fruit. Its attempts at grasping fruit from shrubs succeed with only a 30% probability. The short agent also has a more sensitive stomach, so it can only digest the yellow kind of fruit. The short agent knows which fruits it can digest, but it cannot tell the difference between the height of shrubs or trees (it does not know how far it can reach). The tall agent does not need to interact with the short agent. It can harvest fruit from any tree on its own. The short agent, on the other hand, needs help from the tall agent to get its preferred fruit down from the trees.

Two problems must be overcome if the tall agent is to help the short agent: First, the tall agent must "care" about the well-being of the short agent (i.e. must take it into account in its behavior). Second, the short agent must learn to signal to the tall agent which specific fruits it can digest, since the tall agent cannot perceive that information on its own (see Figure 1).

The environment is either fully mixed where both types of fruit exist alongside each other, or it has a *desert* region where no fruit-bearing plants can grow. When there is a desert, the red fruit only appears in the bottom part of the environment, and the yellow only in the top part. We vary the size of the desert from 0 to denote a fully mixed case to 20 (see Figure 2). We elaborate on this setup in the next section.

4.2 Cost sensitivity moral assessment

We operationalize moral behavior as helping behavior. A helping event occurs when the following sequence of events occur: First, the short agent attempts and fails to grasp a yellow fruit; second, the tall agent then picks it up and drops it; and third, the short agent then picks it up and consumes it.

The short agent may indicate their intention to pick from certain trees, and the tall agent may sacrifice some of their own reward by taking time to pick and hand over yellow fruit to the short agent. The further the taller agent has to deviate from its optimal route to help the small agent, the costlier the helping behavior is.

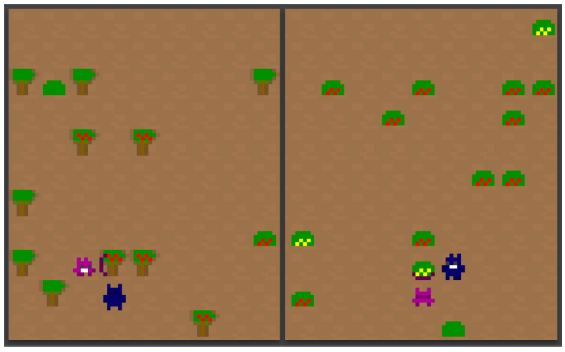

Figure 1: Egocentric views of the same environment, at the same time, for the tall (left) and short (right) agents. Left: The tall agent can distinguish between, and pick, fruit from both trees and shrubs, but all fruits look the same (all fruit looks red). Right: The short agent can distinguish between types of fruit but not the height of trees versus shrubs (all vegetation looks like shrubs), so the agent is unaware of where it can pick from.

We manipulate the distance the tall agent would need to travel in deviation from its self-interested optimal route by varying the distance between the patches (Figure 2). A longer distance means that, if the tall agent decides to help, it must then spend more time in transit. This imposes greater instrumental cost, because it is not possible to eat fruit while traversing the desert.

At evaluation time, the agents will be placed in situations that were not directly part of the training experience. In particular, we reserve some intermediate values of the desert size, as well as some extreme values for evaluation. This way, all behaviors exhibited by the agents (particularly the tall agent) are the result of them adapting at evaluation time to the previously unseen costs. Some test conditions reflect *interpolation*, where the costs are between costs experienced during training; while others are *extrapolation*, where the costs are larger than anything experienced at training time. As discussed before, any change of behavior in the tall agent can be seen as a choice in the presence of new costs resulting from inference of how those costs affect their goals. Care should be taken when interpreting these results to ensure the tall agent hasn't over-fit to their training experience. This is accomplished by comparing their training time behavior (helping or not) with their test time behavior and assessing whether evaluation behavior is congruent with interpolation between training circumstances, and whether the extrapolation is reasonable (e.g., not leading to catastrophic declines in performance).

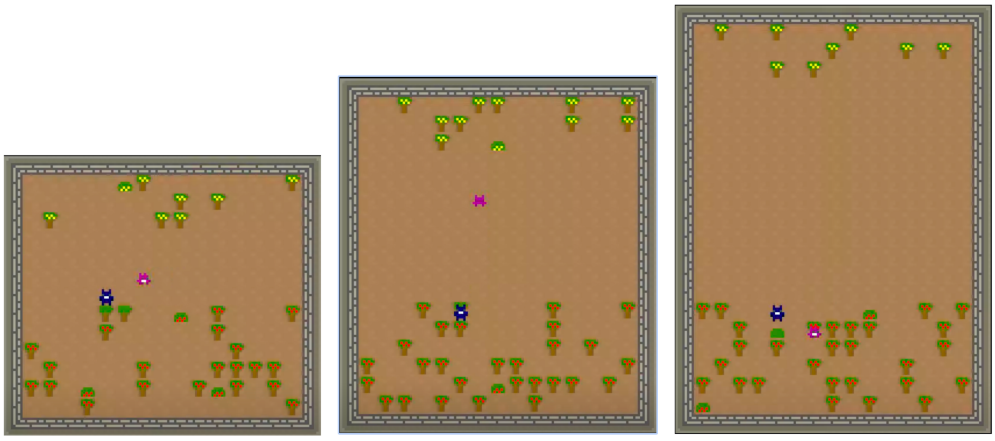

Figure 2: To evaluate the change in helping behavior as cost increases, we increased the distance of fruit-free desert between the yellow fruit patch and the red fruit patch.

4.3 Agents

The agents studied in this report are deep reinforcement learning agents. Agents in the environment have an egocentric partial view window into the environment. This observation window allows agents to see 9 cells ahead, 1 cell behind and 5 cells to each side. Each cell is a region of 8×8 pixels, for a total observations size of 88×88 RGB pixels. The agents learn via the ACB algorithm (see Agapiou et al. (2022); ACB extends prior work on actor-critic RL by Mnih et al. (2016); Espeholt et al. (2018)).

The neural network architecture is comprised of a convolutional network, a feed-forward network, and a memory network from which a policy and a value estimation are produced. The convolutional network has two layers with output channels 16 and 32, respectively. The feed-forward network has two hidden layers of size 64 each. Both the convolutional and feed-forward layers have ReLU activations. The memory is implemented as an LSTM with a hidden size of 256. Both the policy and the value estimation heads are a single layer of size 256.

In addition to the policy gradient loss, the agents have other losses that have become standard in the field. An entropy regularization loss to encourage diversity of policies (with coefficient 0.003). A contrastive-predictive-coding loss applied to the states of the LSTM over time (with coefficient 10.0, and 64 latent space dimensions Oord et al., 2018). A Pop-Art loss to normalize the reward signals (with step size 10^{-3}, and lower and upper bound scales 10^{-2} and 10^{6} respectively following van Hasselt et al., 2016).

Agents are rewarded in two ways: (a) by eating a fruit that is nutritious to them (any fruit for the tall agent, and yellow fruit only for the short agent), and (b) a term related to *advantageous inequity aversion* for the tall agent only, as proposed in Hughes et al. (2018) (following Fehr and Schmidt, 1999). The modified reward of the tall agent is then

$$r'_{tall}(s_t) = r_{tall}(s_t) - \beta \max\left(r_{tall}(s_t) - r_{short}(s_t), 0\right),$$

where $r_{tall}(s_t)$ is the extrinsic environmental reward of the tall agent at time t, and $r_x(s_t) = \lambda r_x(s_{t-1}) + r_x(s_t)$ is the temporally smoothed reward of agent x. That is, they prefer not to let their own rewards too far outstrip those of their partner. We use temporal smoothing with $\lambda = 0.975$ to allow agents to observe the smoothed reward of every player at each timestep. We do not use the disadvantageous inequity aversion term from Hughes et al. (2018); Fehr and Schmidt (1999).

4.4 Training protocol

The training process exposes agents to a variety of cost circumstances associated with the helping behavior (i.e. it was MMRL). As in prior work with such training protocols, policies learnt through this procedure were able to generalize beyond the specific scenarios they experienced during training Wang et al. (2016); Bauer et al. (2023).

At training time, two agents, one tall and one short, are trained in episodes where the environment has a variable desert size of $0, 1, 3, 5, \ldots, 15$. By finding the right θ through a wide range of costs, such that $P_\theta(a, r, s)$ conducts actions(a) that optimize rewards(r) under different costs(s), the agents learn to respond to a particular s at test time.

5 Evaluation and results

We record the number of helping events in an episode using episodes that have a desert of sizes $0, 2, 4, \ldots, 20$ (Note that the range of desert sizes seen during training overlaps with the range seen during evaluation. However, agents never get to see sizes larger than 15 during training). Similar to the volunteer example, for one tall agent, its motivation could be quantified by observing the mean and variance for how often it helps the short agent per episode as a function of the size of the desert.

We train three tall agents with different values of the advantageous inequity coefficient β. We use $\beta = 0.75, 0.5$ and 0.25 for agents 'A', 'B', and 'C,' respectively. We used these values to represent different innate tendencies from the agents to take into account the well-being of others. Alongside the tall agents we train short agents without any extra incentives other than maximization of their selfish reward. We then evaluate these (tall) agents in episodes with the evaluation sizes of the desert as described above.

Recall the two criteria we proposed for comparing moral behavior. One agent can be said to behave more morally than another if it shows:

1. Greater cost insensitivity for morally relevant behaviors; and
2. Adaptive cost sensitivity for morally neutral behaviors.

The tall agents with stronger other-regarding preferences (greater advantageous inequity aversion during training) showed more insensitivity of helping behavior to increasing cost at evaluation time. In Figure 3, Agent A consistently offered more help than Agent B under increasing cost, whereas Agent C did not help at all once costs increased (evaluating criterion 1). We might conclude that Agent A behaved more morally than Agent B, who behaved more morally than Agent C. Critically, however, these agents could merely *appear* cost insensitive for the morally relevant behavior due to general behavioral inflexibility. We would need to also compare their cost responsiveness to morally neutral behaviors (Criterion 2). It would not be a moral case, for example, if the tall agent continues its 'helping' behavior even when the short agent is entirely absent.

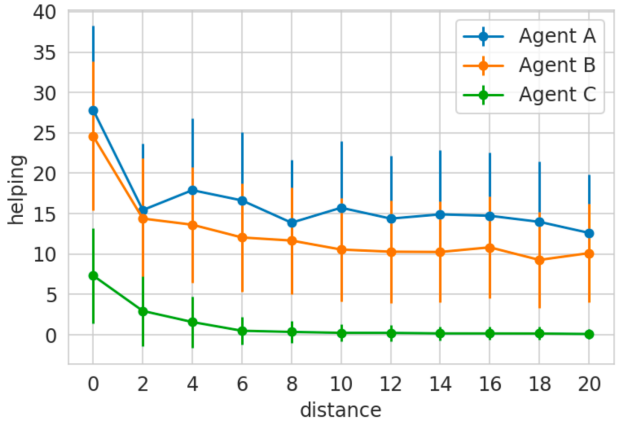

Figure 3: A plot demonstrating how helping behavior of the tall agent decreases as the distance between fruit patches increases (i.e., as cost increases). We evaluate three different agents, each trained with advantageous inequity aversion, which was parameterized as $\beta = 0.75$, 0.5 and 0.25 for agents 'A', 'B', and 'C', respectively.

6 Future Directions

In this work, we proposed a behavior-based cost sensitivity analysis for investigating whether an AI agent is doing the 'right thing for the right reasons,' and we applied this evaluation scheme to a set of deep reinforcement learning agents trained to respond to cost change. With this scheme, it is (in principle) possible to compare human and artificial agent morality in a like-for-like way. As AI capabilities advance and are used in a variety of applications, it will become increasingly important to determine whether agents *merely* do the right thing, or if they do the right thing for the right reasons.

Though we focused in this paper on reinforcement learning agents, we also see opportunity to consider moral evaluation schemes for language models (LM). In the last year, much progress has been made with agents that can leverage human language (e.g., ChatGPT). Pan et al. (2023) demonstrates these agents can be conditioned to score better on a question-answer test of moral choices than reinforcement learning agents. This raises two questions. First, is this evaluation scheme sufficient for comparing the moral cognition of human and artificial agents? "Moral talk" can be cheap, and it is worth considering how to evaluate language models' words in a deeper sense. With the current scheme, it is (in principle) possible to compare human and different artificial agent morality in a like-for-like way. In order to extend our work to LMs, however, we would need to define cost for LMs—which remains an open direction for future research.

Finally, we also see this project as demonstrating the 'virtuous circle' between research in developmental psychology and artificial intelligence (Weidinger et al., 2022). Just as insights from developmental psychology can inform new directions for developing AI (as in the present paper), insights from AI research can similarly shed light on mechanisms of human cognition (e.g., Benton and Lapan, 2022).

References

Agapiou, J. P., Vezhnevets, A. S., Duéñez-Guzmán, E. A., Matyas, J., Mao, Y., Sunehag, P., Köster, R., Madhushani, U., Kopparapu, K., Comanescu, R., et al. (2022). Melting pot 2.0. *arXiv preprint arXiv:2211.13746*.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. (2020). Emergent tool use from multi-agent autocurricula. *International Conference on Learning Representations*.

Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., et al. (2022). Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Bardgett, M. E., Depenbrock, M., Downs, N., Points, M., and Green, L. (2009). Dopamine modulates effort-based decision making in rats. *Behavioral neuroscience*, 123(2):242.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., et al. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, 113(17):4688–4693.

Bauer, J., Baumli, K., Baveja, S., Behbahani, F. M. P., Bhoopchand, A., Bradley-Schmieg, N., Chang, M., Clay, N., Collister, A., Dasagi, V., Gonzalez, L., Gregor, K., Hughes, E., Kashem, S., Loks-Thompson, M., Openshaw, H., Parker-Holder, J., Pathak, S., Nieves, N. P., Rakicevic, N., Rocktäschel, T., Schroecker, Y., Sygnowski, J., Tuyls, K., York, S., Zacherl, A., and Zhang, L. M. (2023). Human-timescale adaptation in an open-ended task space. *ArXiv*, abs/2301.07608.

Benton, D. T. and Lapan, C. (2022). Moral masters or moral apprentices? a connectionist account of sociomoral evaluation in preverbal infants. *Cognitive Development*, 62:101164.

Cousins, M., Atherton, A., Turner, L., and Salamone, J. (1996). Nucleus accumbens dopamine depletions alter relative response allocation in a t-maze cost/benefit task. *Behavioural brain research*, 74(1-2):189–197.

Crosby, M., Beyret, B., and Halina, M. (2019). The animal-AI olympics. *Nature Machine Intelligence*, 1(5):257–257.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2):353–380.

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, 6:97–103.

Cushman, F. and Mele, A. (2008). Intentional action. *Experimental philosophy*, 171.

Cushman, F., Sheketoff, R., Wharton, S., and Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, 127(1):6–21.

Cushman, F. and Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive science*, 35(6):1052–1075.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. (2018). Impala: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *International conference on machine learning*, pages 1407–1416. PMLR.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108.

Haas, J. (2020). Moral gridworlds: A theoretical proposal for modeling artificial moral cognition. *Minds and Machines*, 30(2):219–246.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Hamlin, J., Wynn, K., and Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental science*, 13(6):923–929.

Hamlin, J. K., Wynn, K., and Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450(7169):557–559.

Hamlin, J. K., Wynn, K., Bloom, P., and Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the national academy of sciences*, 108(50):19931–19936.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., et al. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31.

Kant, I. and Schneewind, J. B. (2002). *Groundwork for the Metaphysics of Morals*. Yale University Press.

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in cognitive sciences*, 9(8):357–359.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and brain sciences*, 33(4):315–329.

Köster, R., Hadfield-Menell, D., Everett, R., Weidinger, L., Hadfield, G. K., and Leibo, J. Z. (2022). Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences*, 119(3):e2106028118.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Leibo, J. Z., d'Autume, C. d. M., Zoran, D., Amos, D., Beattie, C., Anderson, K., Castañeda, A. G., Sanchez, M., Green, S., Gruslys, A., et al. (2018). Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*.

Leibo, J. Z., Dueñez-Guzman, E. A., Vezhnevets, A., Agapiou, J. P., Sunehag, P., Koster, R., Matyas, J., Beattie, C., Mordatch, I., and Graepel, T. (2021). Scalable evaluation of multi-agent reinforcement learning with Melting Pot. In *International conference on machine learning*, pages 6187–6199. PMLR.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Markovits, J. (2010). Acting for the right reasons. *Philosophical Review*, 119(2):201–242.

Marshall, J. and McAuliffe, K. (2022). Children as assessors and agents of third-party punishment. *Nature Reviews Psychology*, 1(6):334–344.

Marshall, J., Yudkin, D. A., and Crockett, M. J. (2021). Children punish third parties to satisfy both consequentialist and retributive motives. *Nature Human Behaviour*, 5(3):361–368.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences*, 11(4):143–152.

Mikulik, V., Delétang, G., McGrath, T., Genewein, T., Martic, M., Legg, S., and Ortega, P. (2020). Meta-trained agents implement bayes-optimal agents. *Advances in neural information processing systems*, 33:18691–18703.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR.

Monroe, A. E. and Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General*, 146(1):123.

Newman, G. E. and Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological science*, 25(3):648–655.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pan, A., Shern, C. J., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., and Hendrycks, D. (2023). Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *arXiv preprint arXiv:2304.03279*.

Rai, T. S. and Holyoak, K. J. (2010). Moral principles or consumer preferences? alternative framings of the trolley problem. *Cognitive Science*, 34(2):311–321.

Railton, P. (2017). Moral learning: Conceptual foundations and normative relevance. *Cognition*, 167:172–190.

Randall, P. A., Pardo, M., Nunes, E. J., López Cruz, L., Vemuri, V. K., Makriyannis, A., Baqi, Y., Müller, C. E., Correa, M., and Salamone, J. D. (2012). Dopaminergic modulation of effort-related choice behavior as assessed by a progressive ratio chow feeding choice task: pharmacological studies and the role of individual differences.

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., and Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of cognitive neuroscience*, 18(5):803–817.

Shenhav, A. and Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, 67(4):667–677.

Singh, S. P., Barto, A. G., and Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tasimi, A. and Wynn, K. (2016). Costly rejection of wrongdoers by infants and children. *Cognition*, 151:76–79.

van Hasselt, H. P., Guez, A., Hessel, M., Mnih, V., and Silver, D. (2016). Learning values across many orders of magnitude. *Advances in neural information processing systems*, 29.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860–868.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Warneken, F., Hare, B., Melis, A. P., Hanus, D., and Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS biology*, 5(7):e184.

Warneken, F. and Tomasello, M. (2009). The roots of human altruism. *British Journal of Psychology*, 100(3):455–471.

Weidinger, L., Reinecke, M. G., and Haas, J. (2022). Artificial moral cognition: Learning from developmental psychology. *PsyArXiv*.

Woo, B. M. and Spelke, E. S. (2023). Toddlers' social evaluations of agents who act on false beliefs. *Developmental Science*, 26(2):e13314.

Woo, B. M., Tan, E., and Hamlin, J. K. (2022). Human morality is based on an early-emerging moral core. *Annual Review of Developmental Psychology*, 4:41–61.

Young, L. and Dungan, J. (2012). Where in the brain is morality? everywhere and maybe nowhere. *Social neuroscience*, 7(1):1–10.

Young, L. and Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and personality psychology compass*, 7(8):585–604.

Yudkin, D. A., Van Bavel, J. J., and Rhodes, M. (2020). Young children police group members at personal cost. *Journal of Experimental Psychology: General*, 149(1):182.